CIDU
CONFERENCE ON
INTELLIGENT DATA UNDERSTANDING

Anomaly Detection in Massive Data Streams and Archives

Eamonn Keogh
Computer Science & Engineering Department
University of California - Riverside
Riverside, CA 92521

Motifs and Discords are all you need to find anomalies and regularities in time series data

Controversial claim!

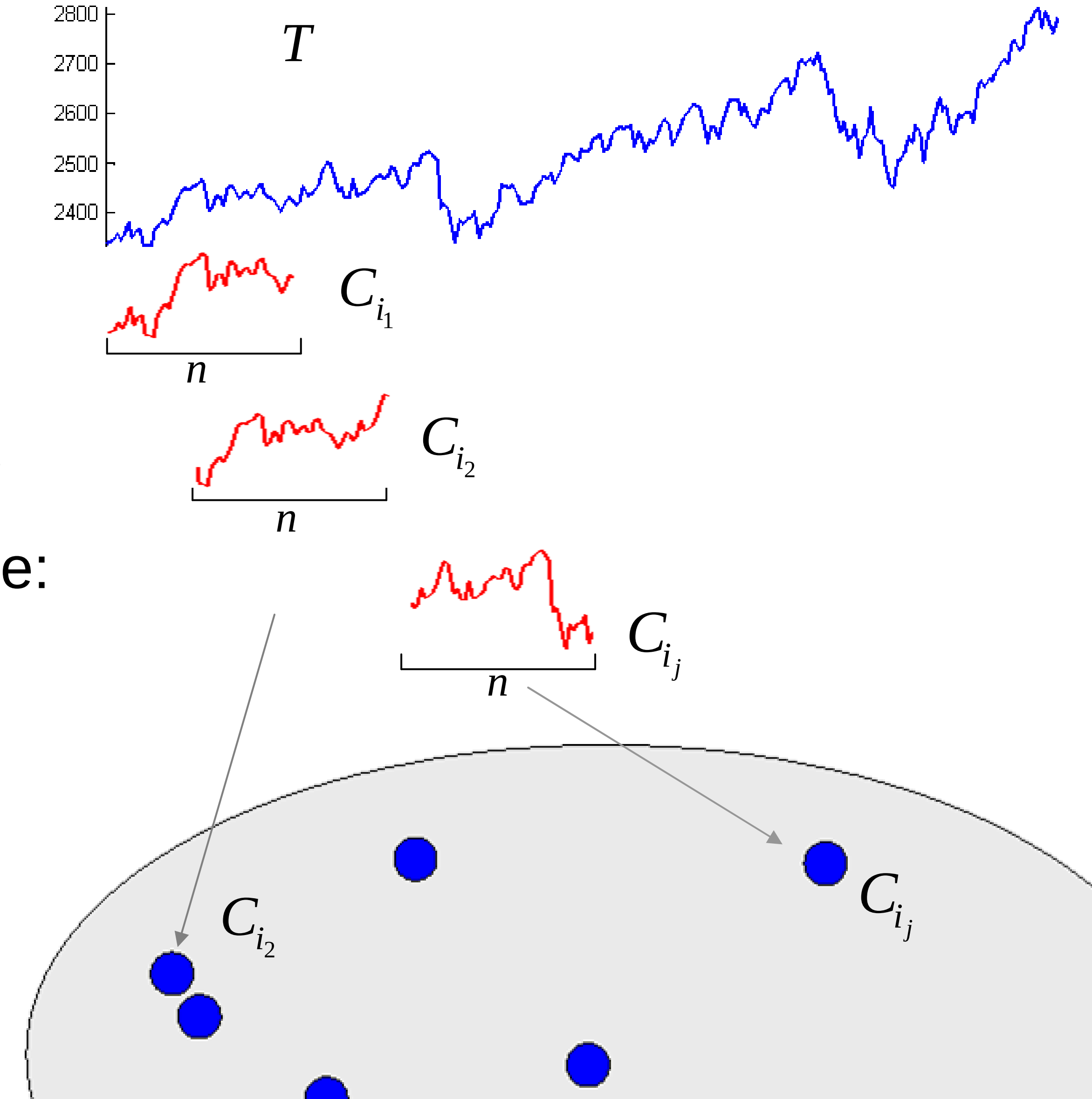
Outline

- What are motifs/discords?
  - What are motifs joins?
- How can we find them efficiently (gloss over)
- How can I use them to mine my data and to detect anomalies?
- Do they really work? (lots of case studies)
- Conclusions/open problems

Definitions and assumptions

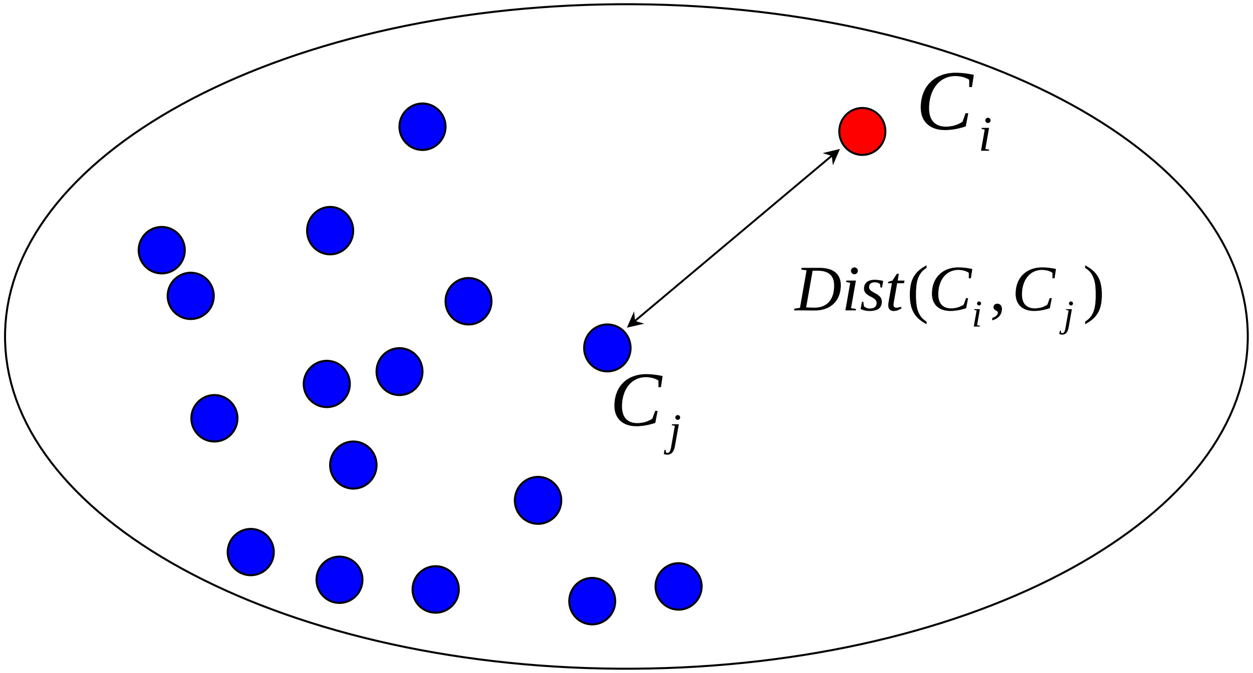
- Notation
  - time series:
    $T = (t_1, \dots, t_m)$
  - subseqence:
    $C_i = (t_p, \dots, t_{p+n-1})$
    $n \leq m,\ 1 \leq p \leq m - n + 1$
  - time series database:
    $S = \{C_i\}$

Function $Dist(C_i, C_j)$ (lets assume Euclidean distance) defines an ordering for the elements in S



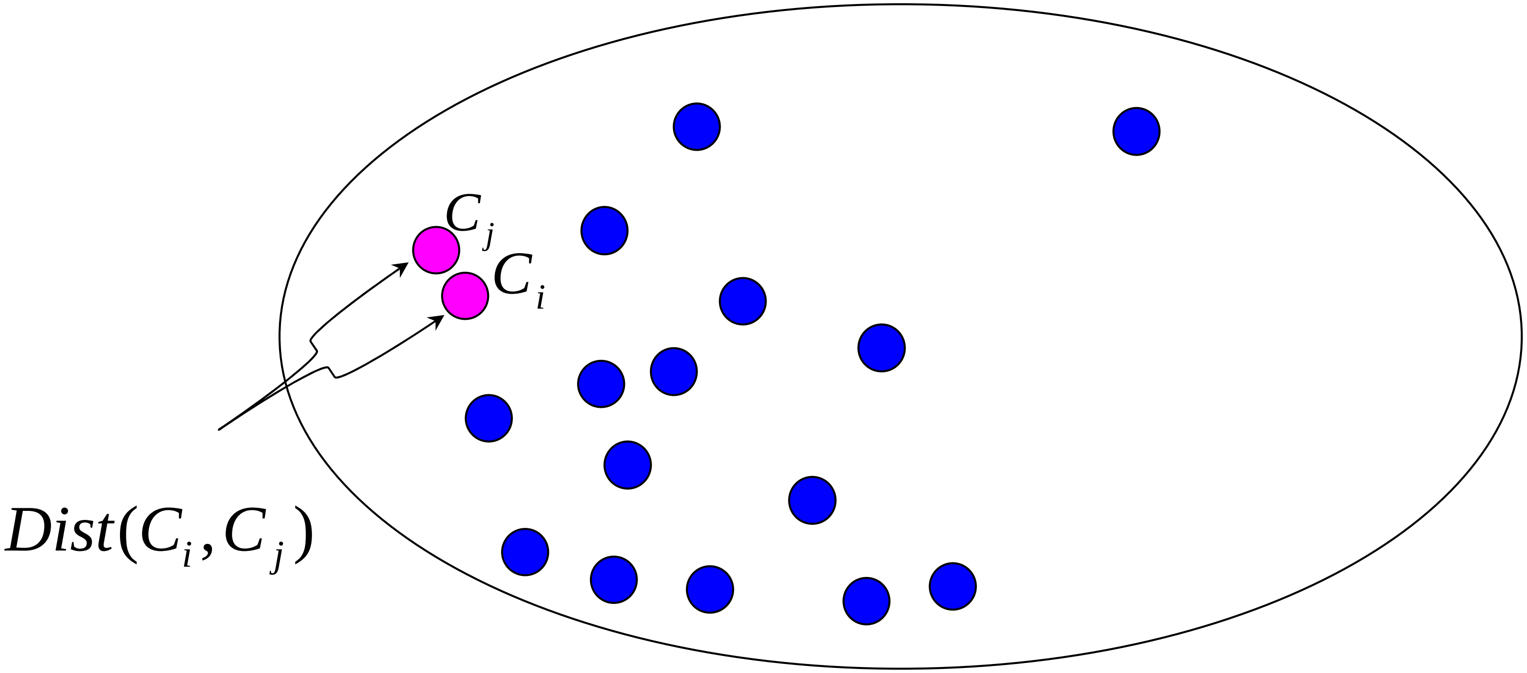
Time series discords

- *Most-significant discord* – the subsequence $C_i \in S$ with maximal distance $Dist(C_i, C_j)$ to its nearest neighbor $C_j \in S$



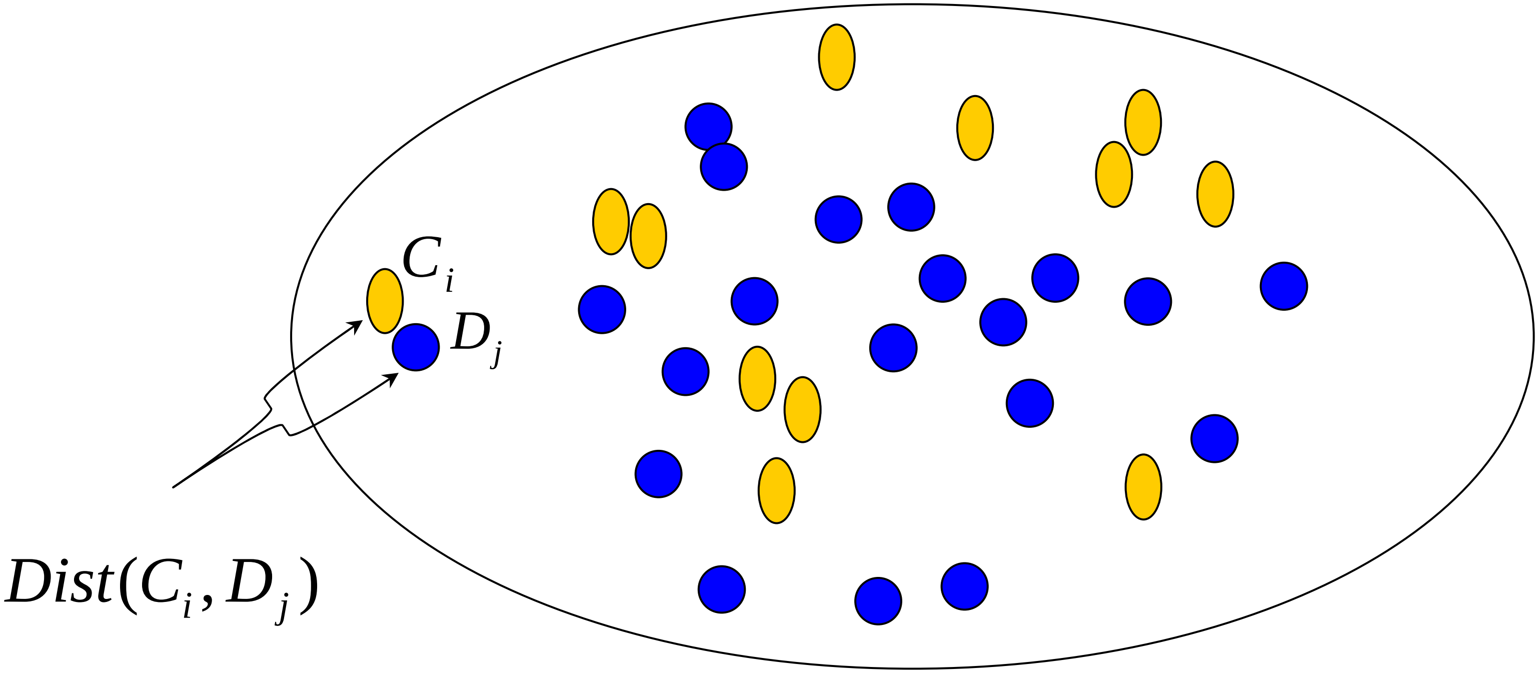
Time series motifs

- *Time Series Motifs* – the pair of subsequences $C_i, C_j \in S$ with minimal distance to each other



Motifs Joins

- *Motif Joins* – the pair of subsequences, *one of each color*, $C_i, D_j \in S$ with minimal distance to each other

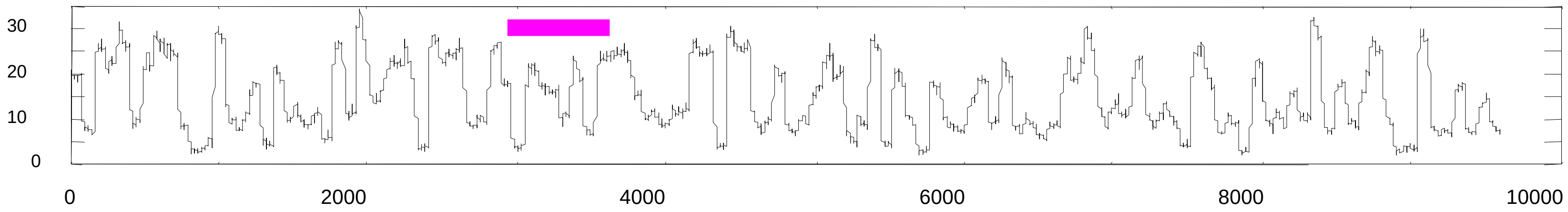


Motifs and Discords

- The definitions are slightly more complex, in order to eliminate pathological cases.
- The naïve algorithm to find motifs or discords is quadratic in time.
- Some of the talk “Finding Repetitive Sequential Patterns for Discovering Anomalies”, presented here may be related to motifs

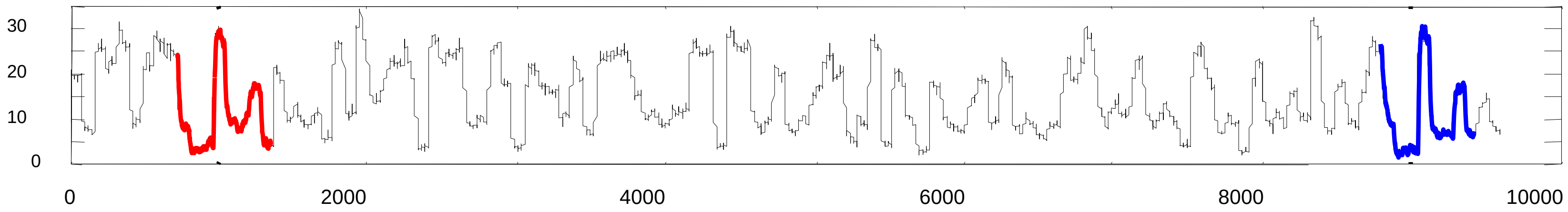
What are Time Series Motifs?

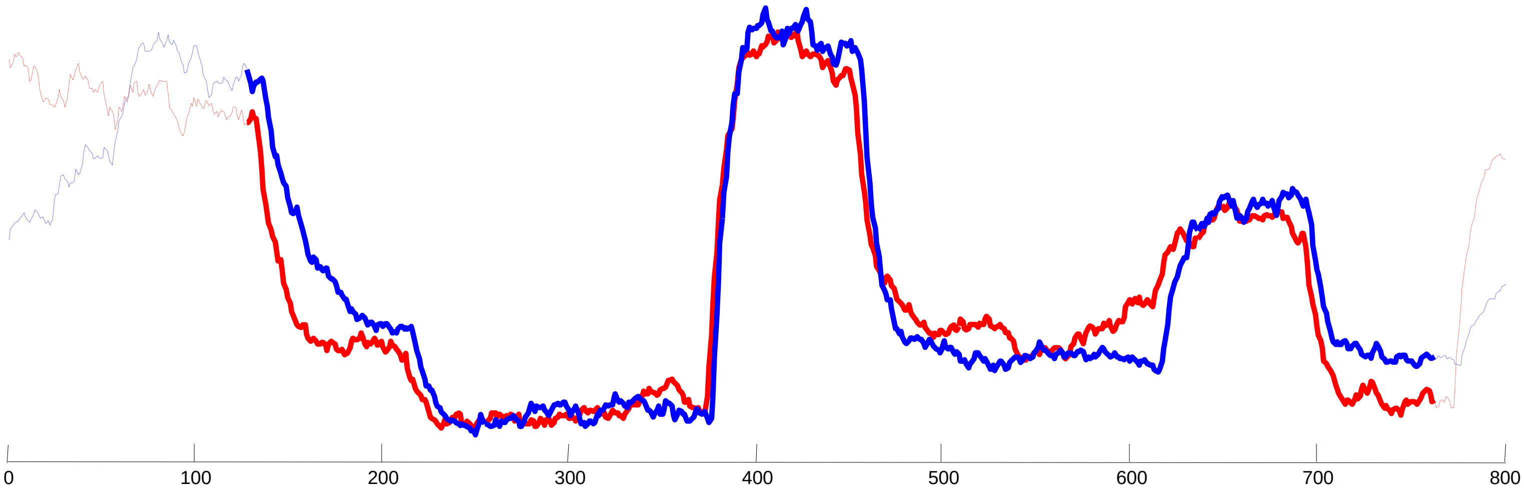
Industrial Steam Generator Dataset



Are there any repeated subsequences of this length ▬ in the sequence above?

What are Time Series Motifs?



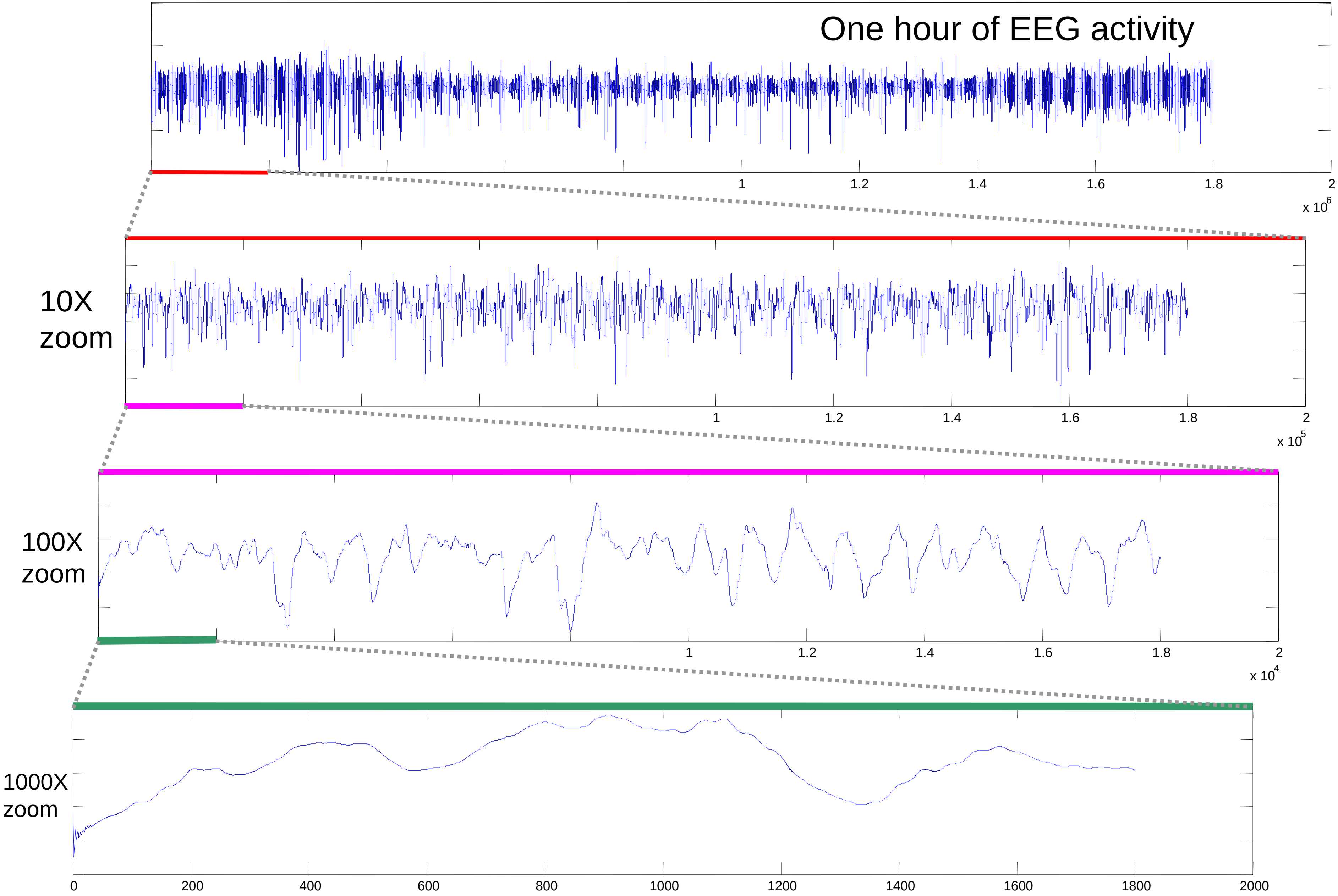


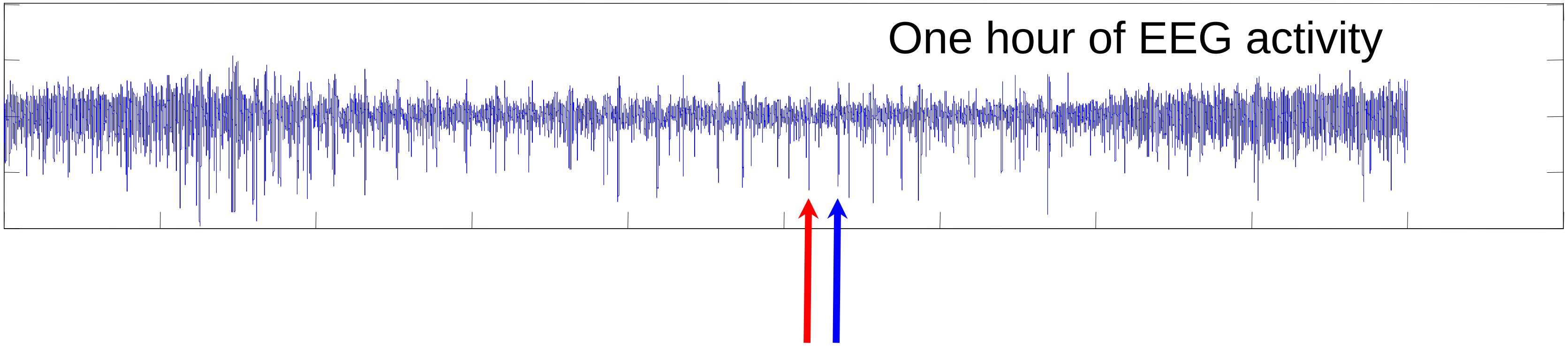
A case study in motif discovery

A month ago I was approached by two Ph.d/MD from the Harvard health system, working in Brigham and Women's Hospital and Massachusetts General Hospital (MGH).

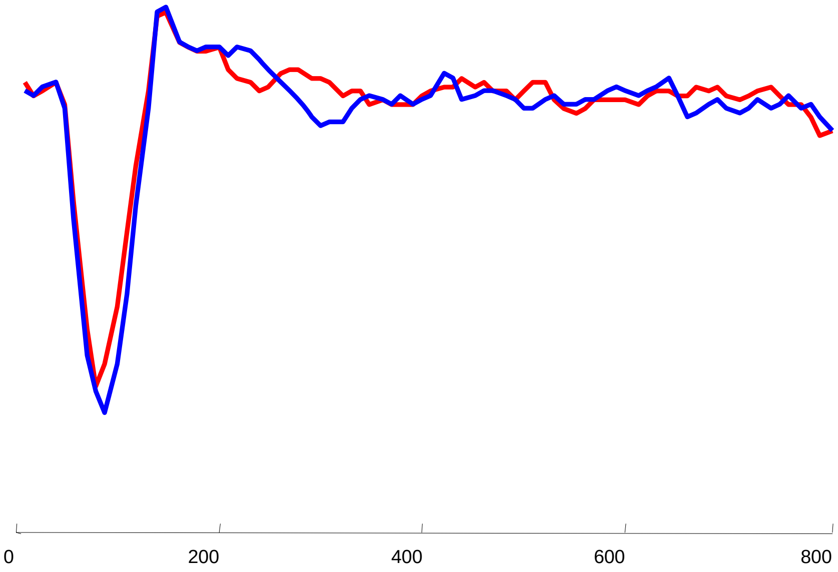
They want to build a “dictionary” of all EEG patterns

However, to process just one hour of EEG data was taking them 24 hours…


One hour of EEG activity
1
1.2
1.4
1.6
1.8
2
x 10^6
10X zoom
1
1.2
1.4
1.6
1.8
2
x 10^5
100X zoom
1
1.2
1.4
1.6
1.8
2
x 10^4
1000X zoom
0
200
400
600
800
1000
1200
1400
1600
1800
2000



Of the approximately five billion possible pairings of subsequences of length 800, taken from the EEG trace above, this pair has the smallest Euclidean distance



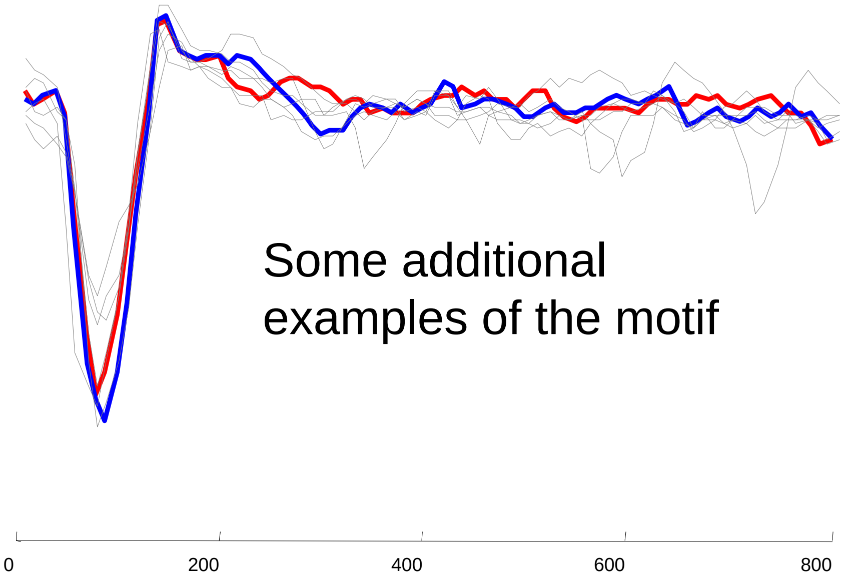
Sanity Check Questions I

Is this a coincidence?

Would we find two equally similar patterns in random data?

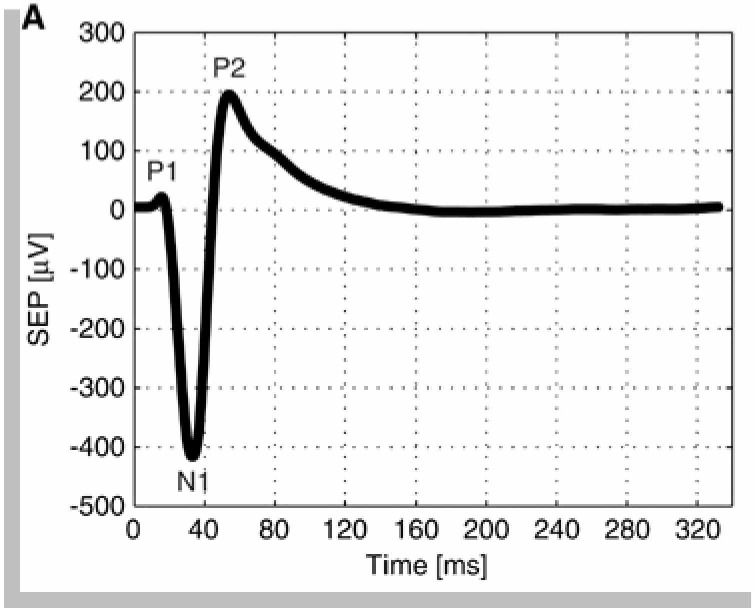
Informal Answer: There really does seem to be a lot of similar patterns.

For the moment we are working on the assumption that the pattern must be conserved for some reason

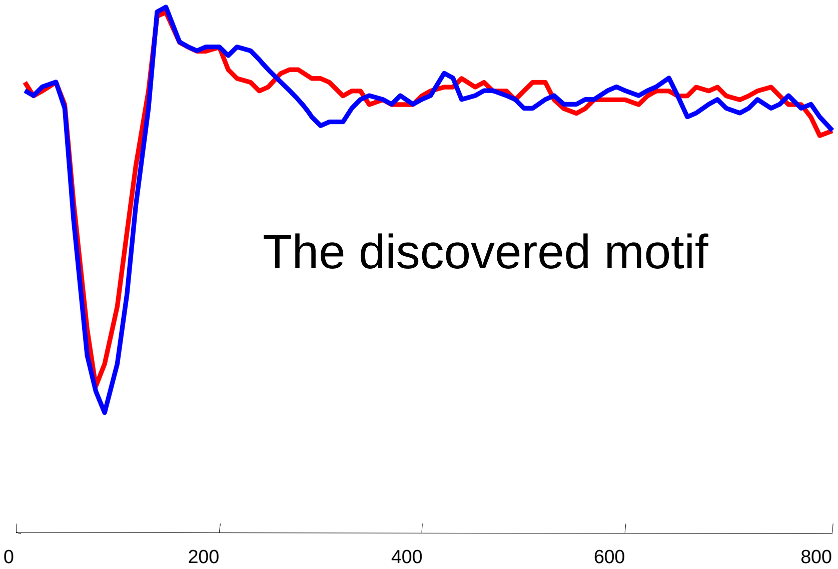

Some additional examples of the motif

Sanity Check Questions II

Does the pattern have any biological meaning?



B. J. Stefanovic, W. Schwindt, M. Hoehn and A. C. Silva, Functional uncoupling of hemodynamic from neuronal response by inhibition of neuronal nitric oxide synthase, Journal of Cerebral Blood Flow & Metabolism, 27: 741–754, 2007.



The discovered motif

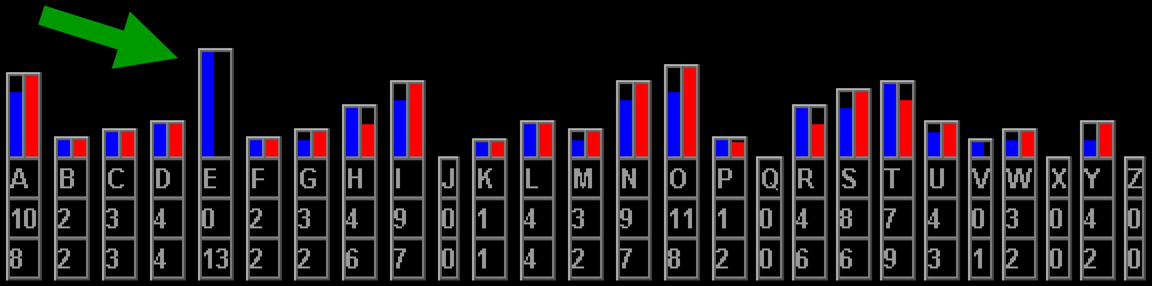
…upon this basis I am going to show you how a bunch of bright young folks did find a champion; a man with boys and girls of his own; a man of so dominating and happy individuality that Youth is drawn to him as is a fly to a sugar bowl. It is a story about a small town. It is not a gossipy yarn; nor is it a dry, monotonous account, full of such customary "fill-ins" as "romantic moonlight casting murky shadows down a long, winding country road.". In contrast Eamonn has been known to study of ozone on zebras' laziness in Zaire and Zanzibar, a fact that has endeared him to the emperor of ….

What use are these motifs?

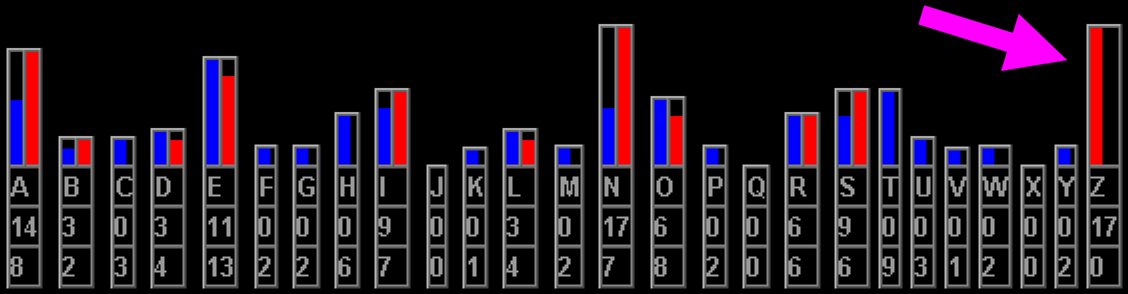
Consider the stream of text above, suppose we want to detect unusual regions in it. We could begin by comparing the observed frequencies of letters to the predicted frequencies of letters….

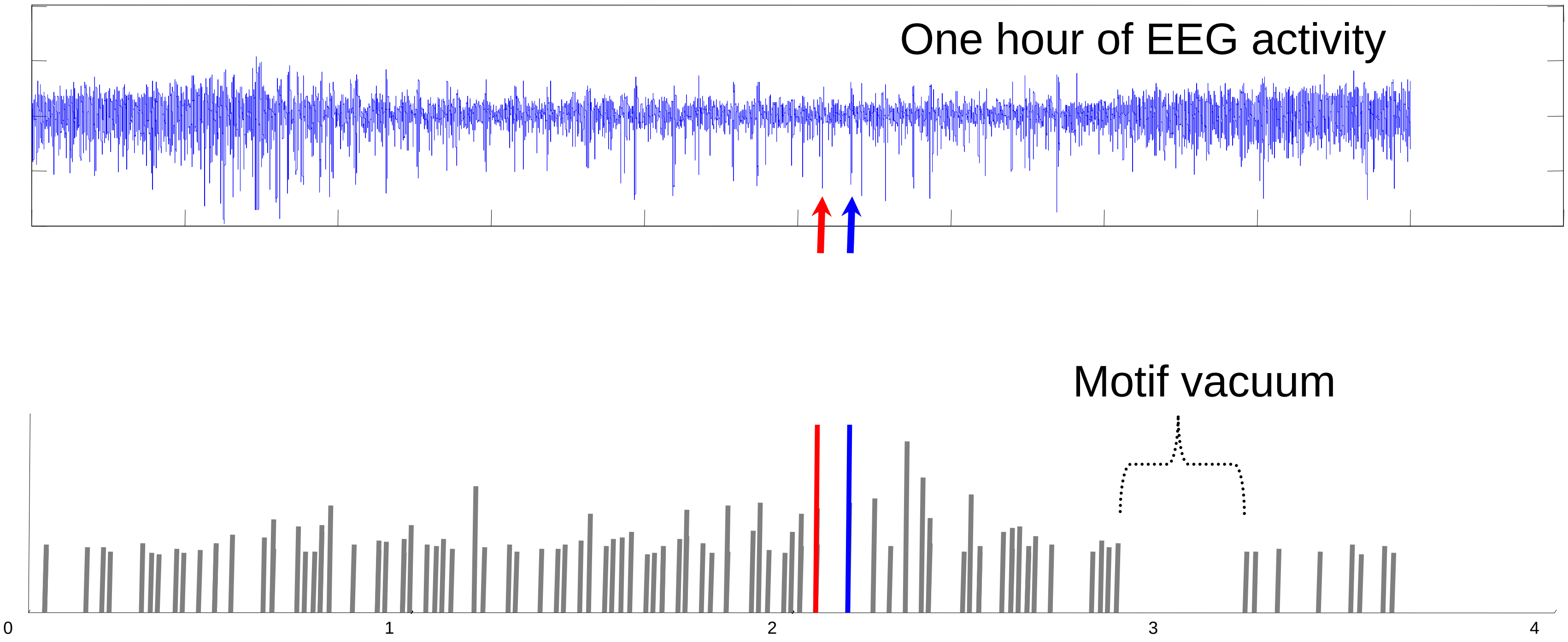
…upon this basis I am going to show you how a bunch of bright young folks did find a champion; a man with boys and girls of his own; a man of so dominating and happy individuality that Youth is drawn to him as is a fly to a sugar bowl. It is a story about a small town. It is not a gossipy yarn; nor is it a dry, monotonous account, full of such customary "fill-ins" as "romantic moonlight casting murky shadows down a long, winding country road.". In contrast Eamonn has been known to study of ozone on zebras' laziness in Zaire and Zanzibar, a fact that has endeared him to the emperor of ….

In the green region the letter ‘E’ is vastly underrepresented



In the pink region the letter ‘Z’ is vastly overrepresented





Here we plotted the locations of the 90 occurrences of the motif (the taller the marker, the closers they are to the red/blue prototypes).

For the most part, the motifs are uniformly distributed.

However there appears to be a region where there is a relative dearth of motif occurrences. Does this mean anything?

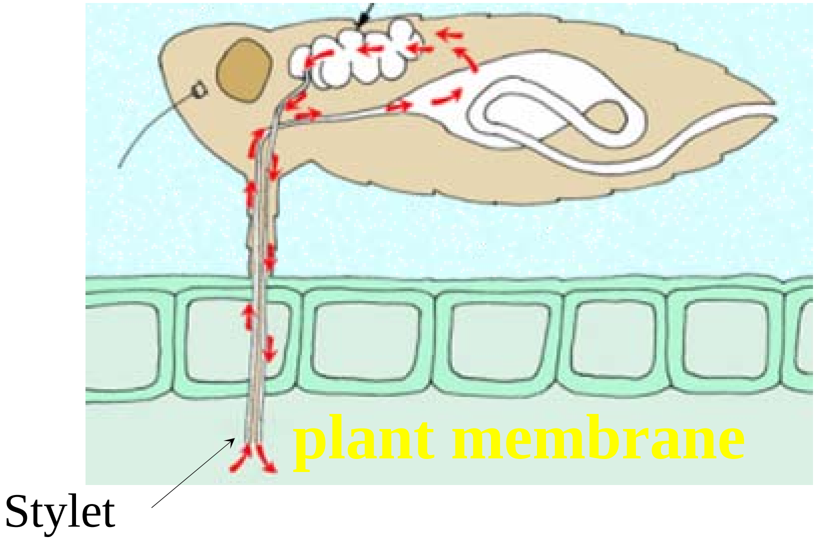
Another Case Study

Beet Leafhopper
(*Circulifer tenellus*)

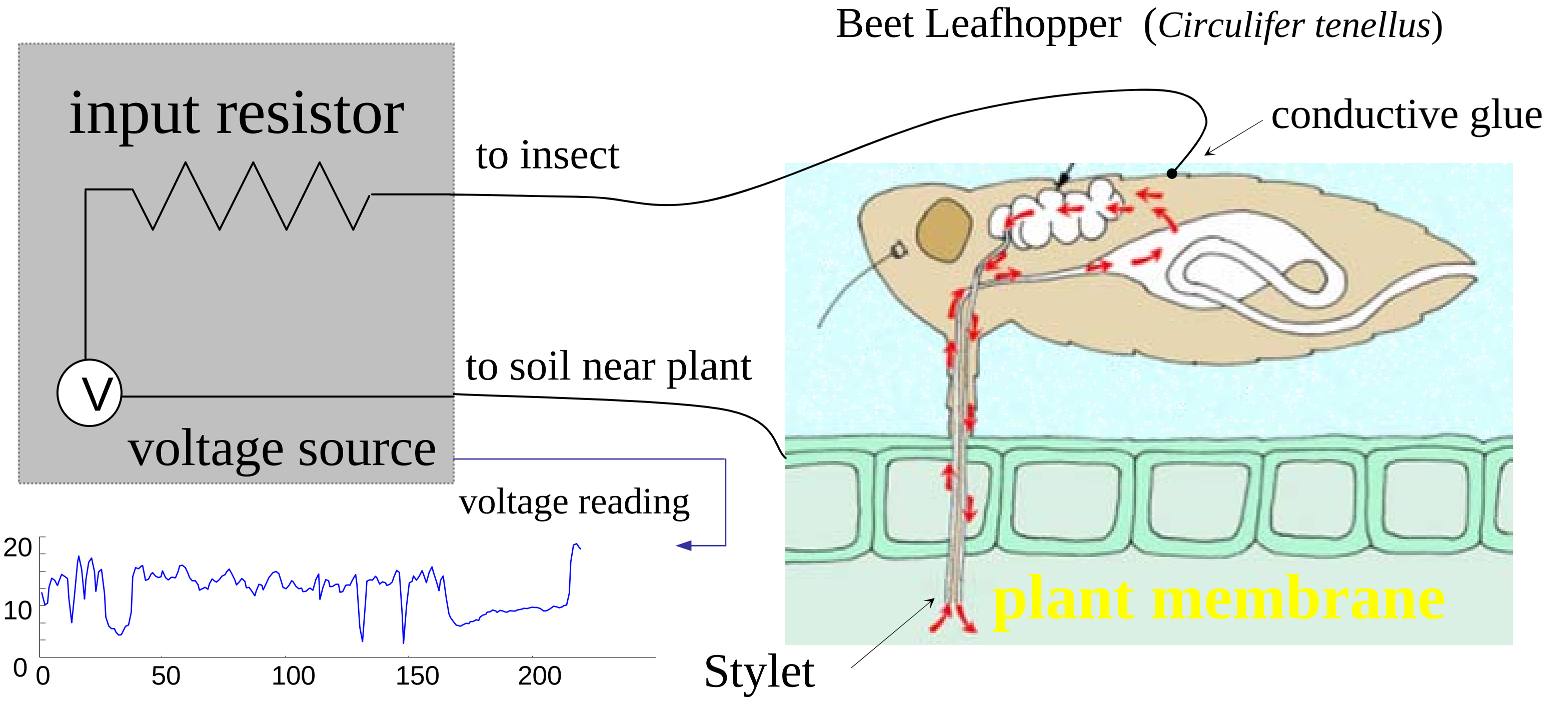
Photo by A.C. Magyarosy, Bugwood.org

In North America, the Beet Leafhopper is the only known vector (carrier) of curly top virus, which causes major economic losses in a number of crops including sugarbeet, tomato, and beans.

Beet Leafhopper (*Circulifer tenellus*)

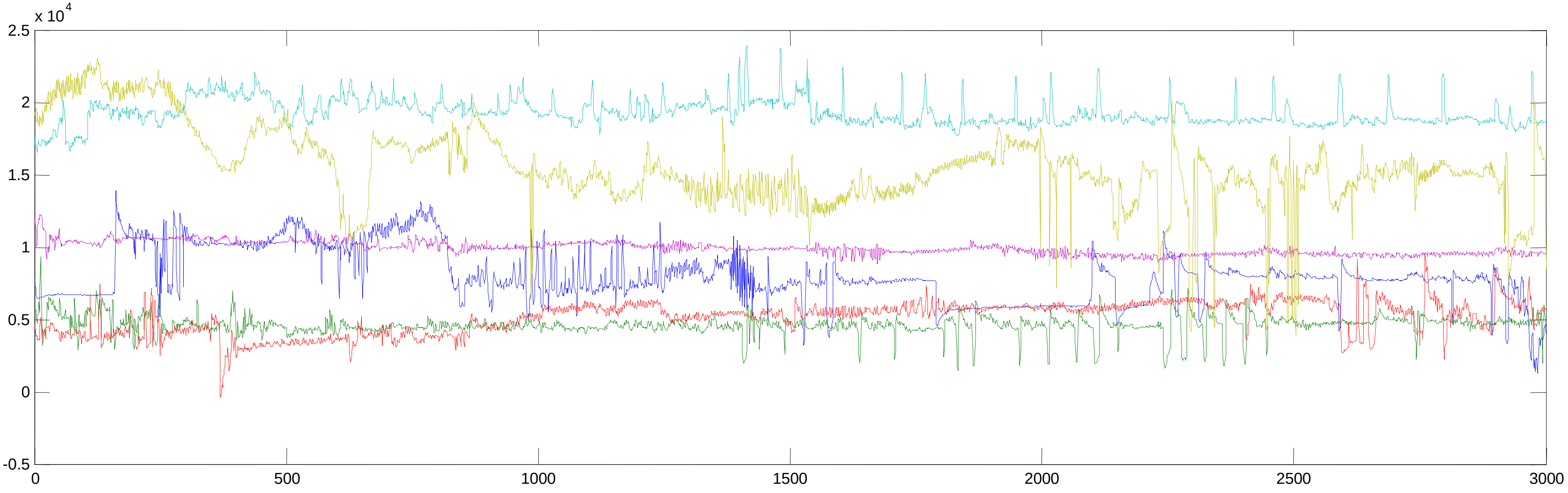


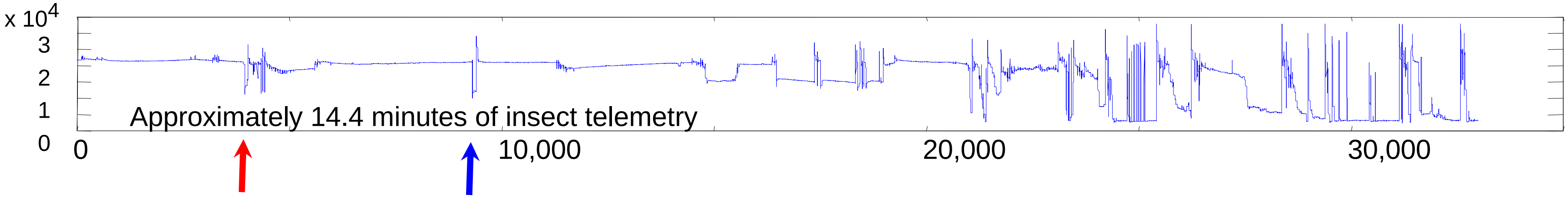
Good News: We can wire up the insect, and we have telemetry!



Bad News: The data is messy

Below are five traces of length 3,000, but we have *thousands* of traces with lengths in the *millions*. How can we make sense of this data?

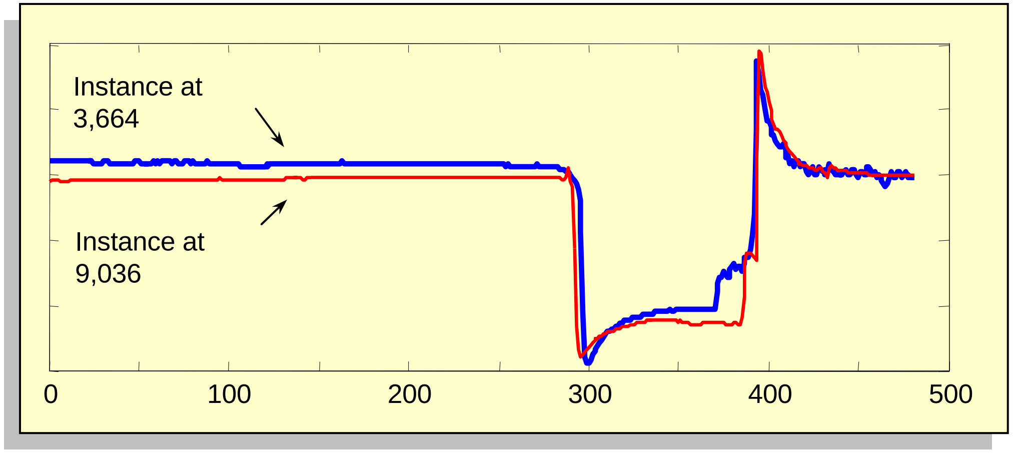




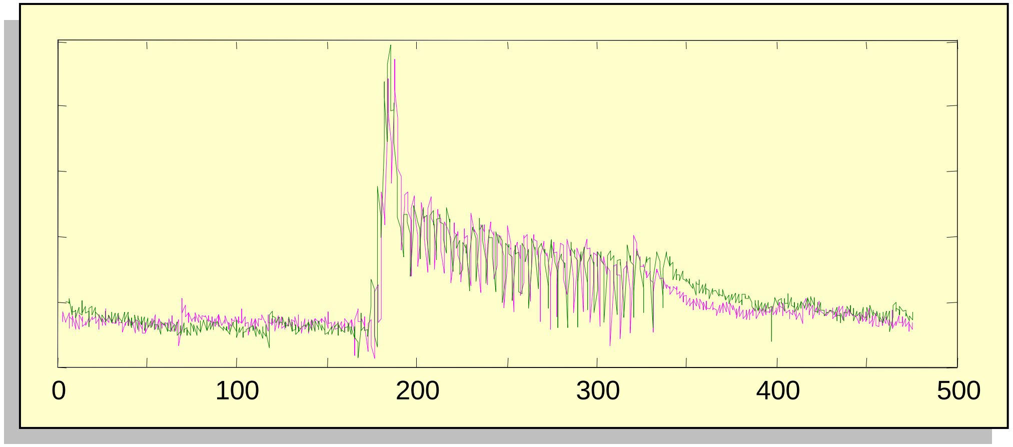
…how can we make sense of this data? We can start by finding motifs.

- Here is a startling motif found in a 14 minute subsection of a longer trace

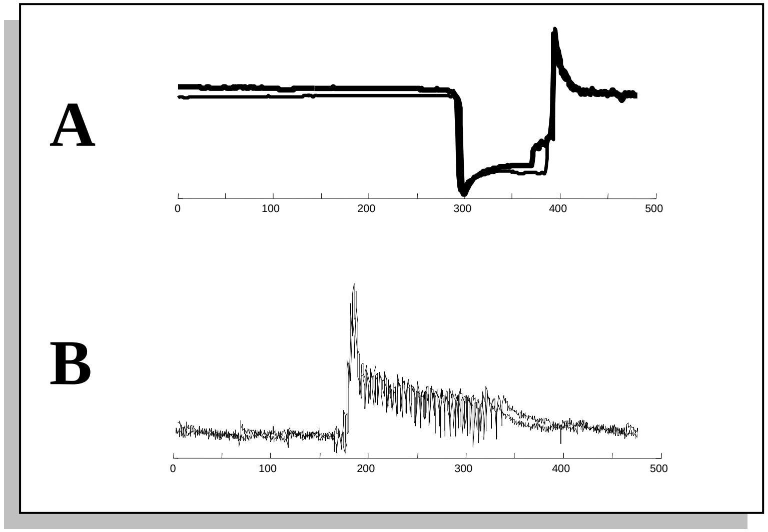
Zoom in factor 30X



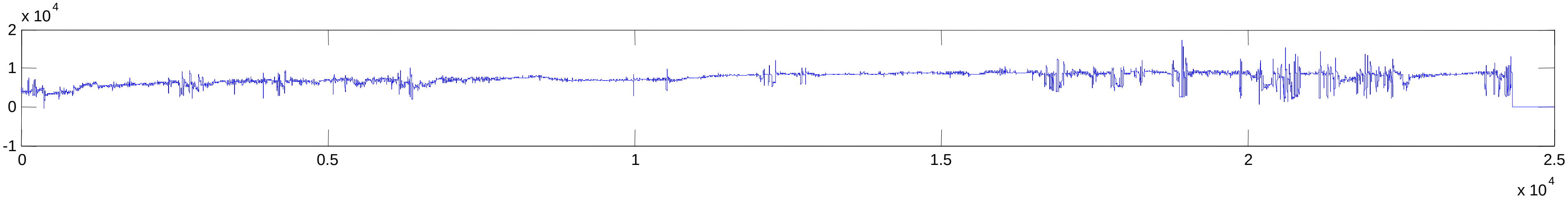
- And here is the second motif



We can give the motifs discrete labels, **A** , **B**, **C** etc



Then a long, noisy, high dimensional time series like this…



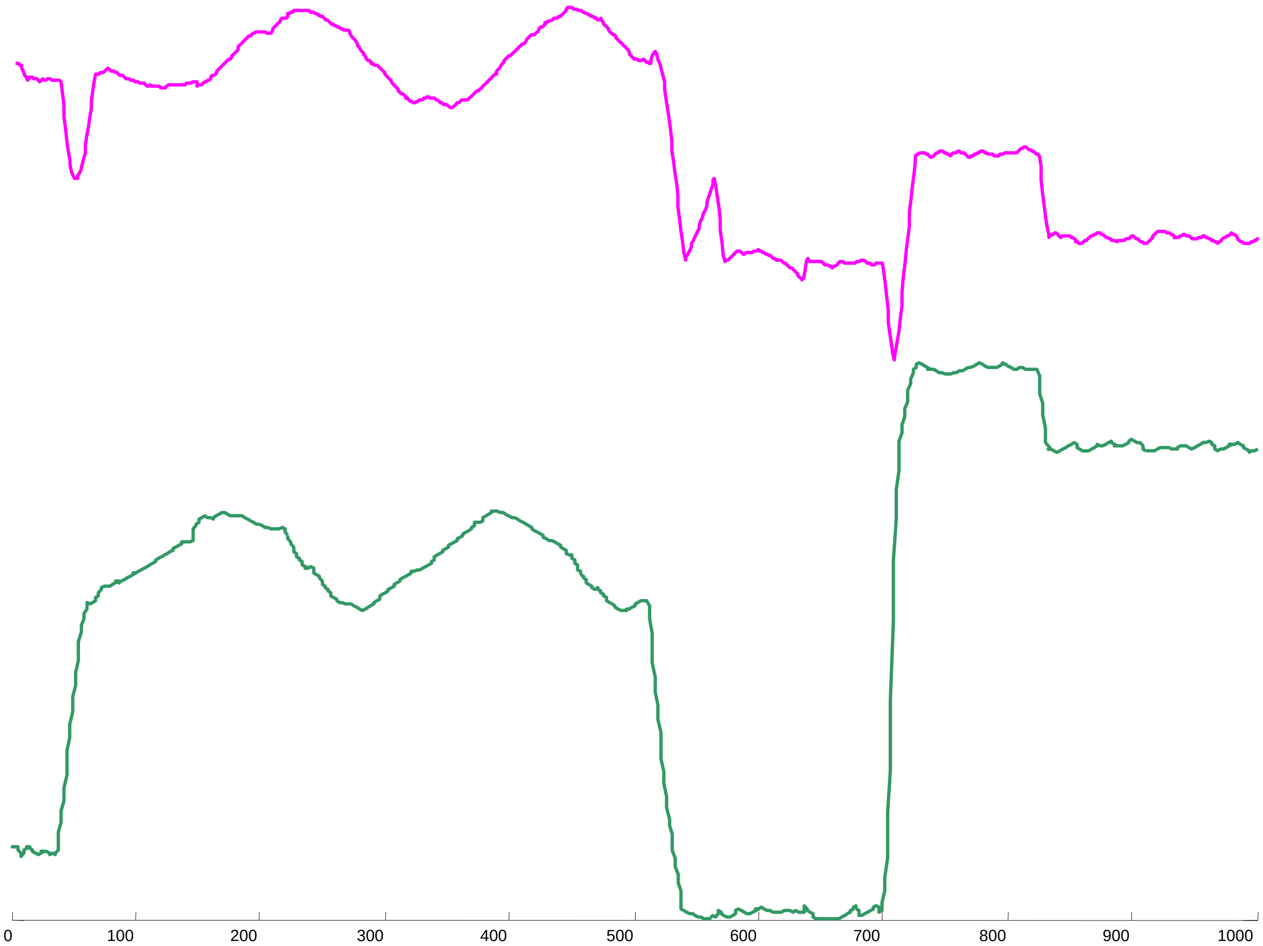
Becomes a simple string like this…

B B C A B B C A C A B A B B C C C B B C A B A B B A C…

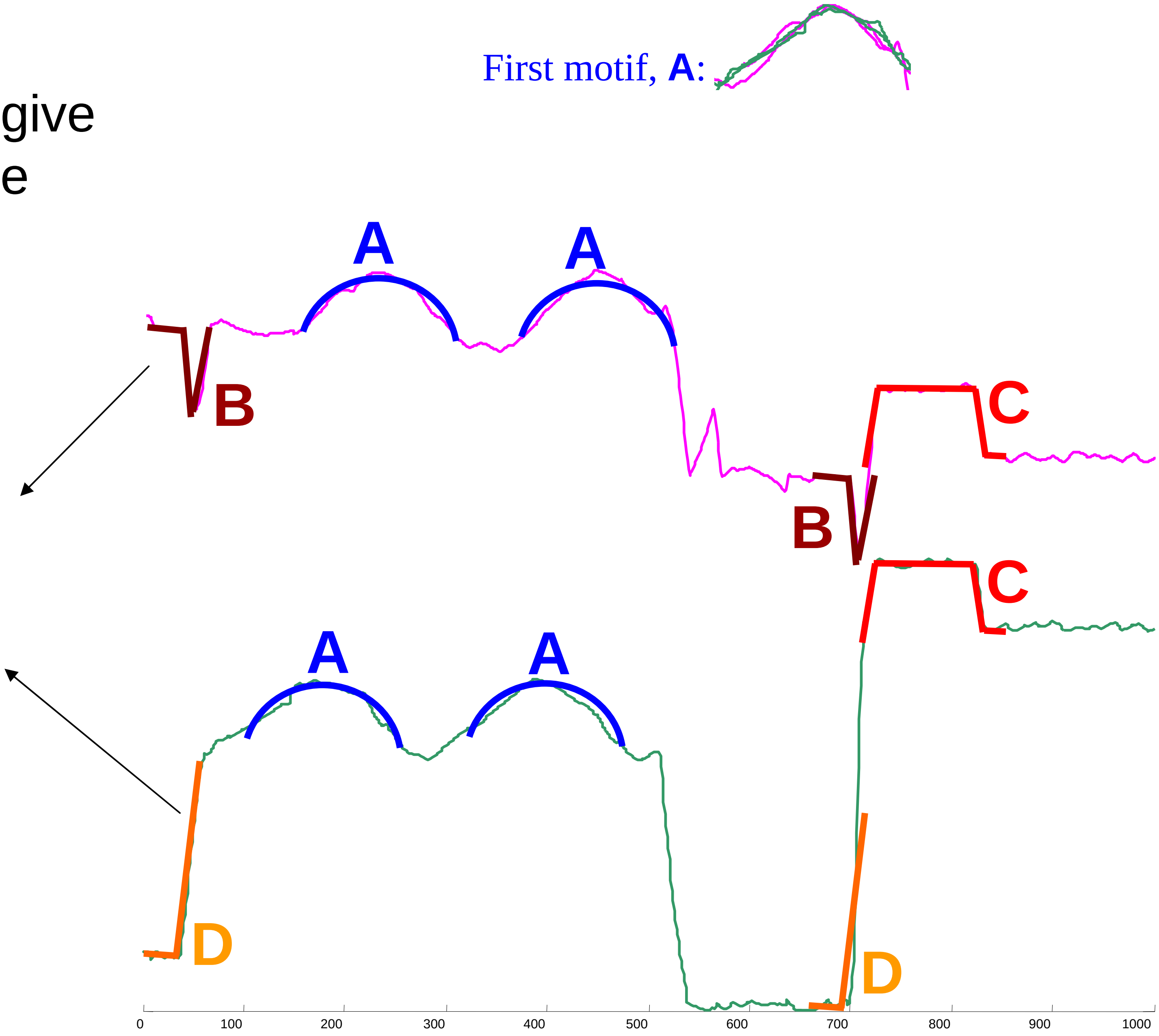
And then we can mine **rules** like this…

If you see ‘**B**’ followed by ‘**B**’, then the next symbol will be ‘**C**’, with a 80% probability.

Space Shuttle
STS-57 Telemetry



Find motifs, give them discrete labels…



B A A B C

D A A D C

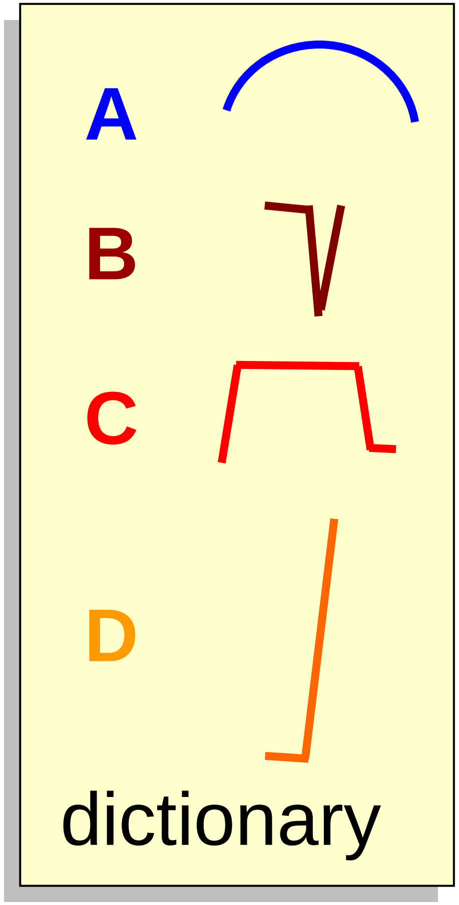
We can now use off the shelf string/DNA algorithms to look for patterns…

B A A B C

D A A D C

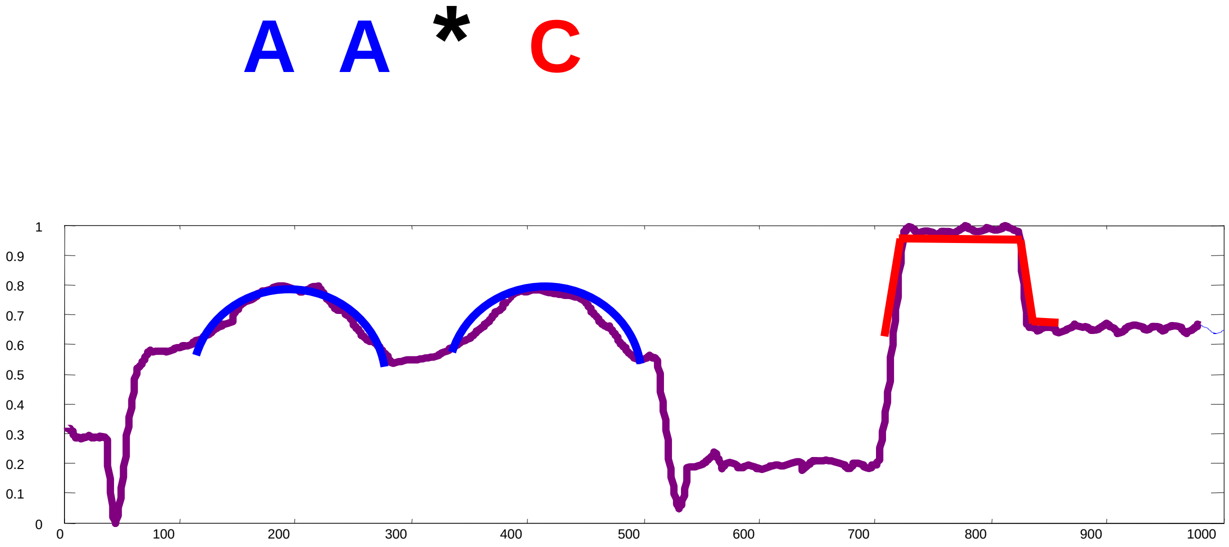
If we use a “don’t-care” we find the pattern….

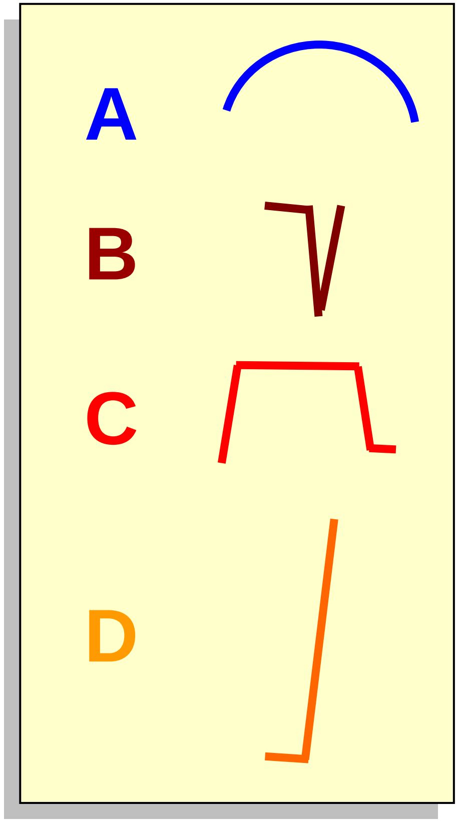
A A * C



We can now monitor streaming data for this pattern. In particular, if we see the prefix **A A**, then we can expect to see pattern **C** within 11 minutes.

In our contrived example, this really works!



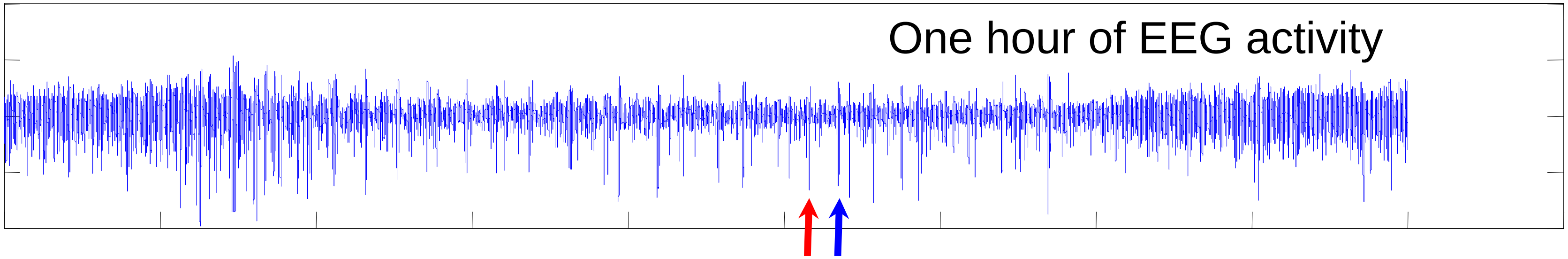


Monitoring for the dictionary of patterns can be done efficiently with *Atomic Wedgie*

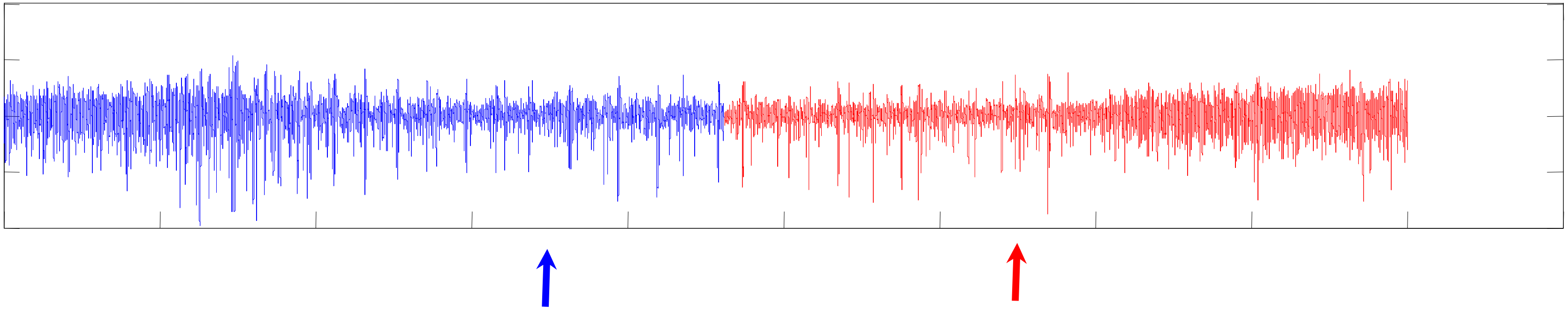
L. Wei, E. Keogh, H. Van Herle, and A. Mafra-Neto (2005). Atomic Wedgie: Efficient Query Filtering for Streaming Time Series. ICDM 2005

Motifs Joins

Recall our definition of motifs….

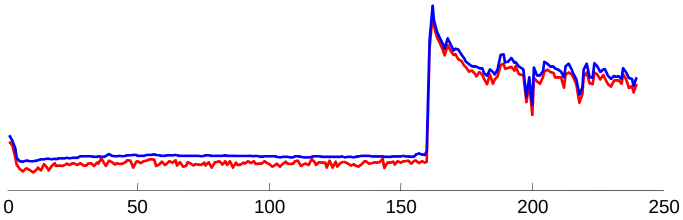


Suppose we augment it by saying that one of the pair must come from set **A**, and one from set **B**.

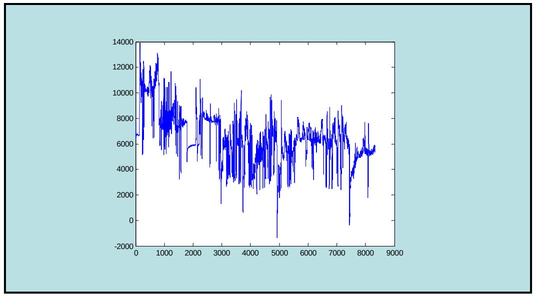


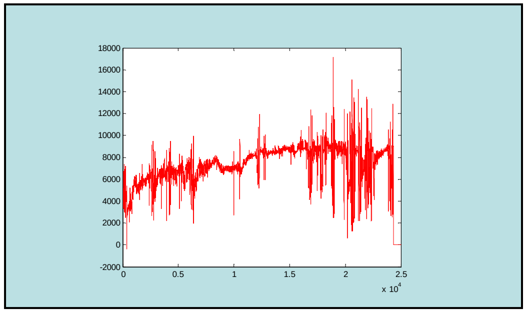
Motifs Joins II

We could do a motif join between the telemetry from the rocket that blew this year, and the rocket that blew up five years ago…



This year



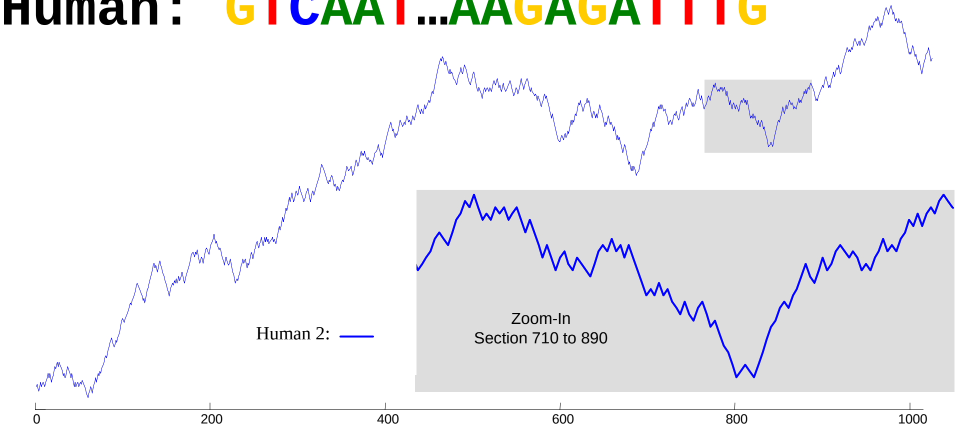


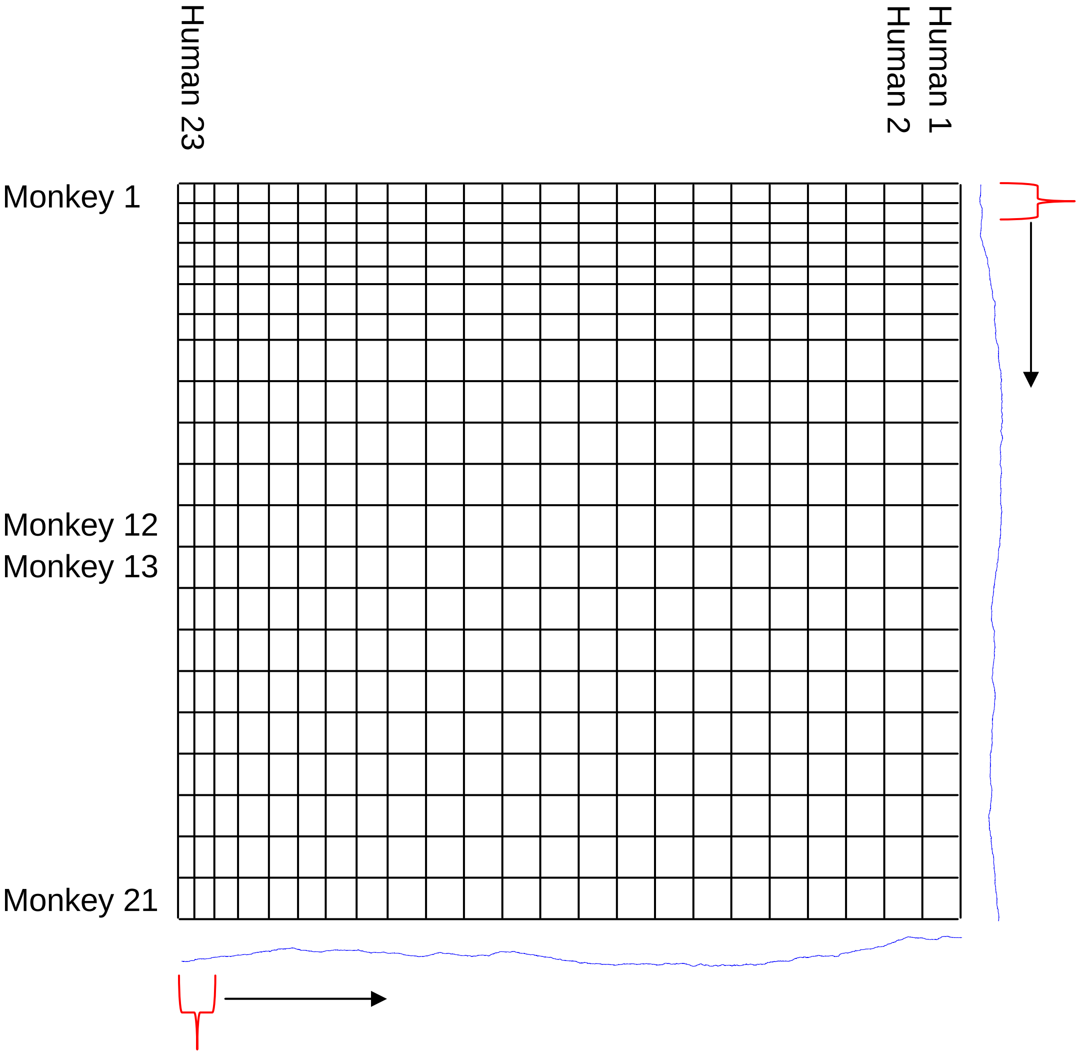
Five years ago

An algorithm to convert DNA strings to time series

```
$T_1 = 0$;
For i = 1 to length(DNAstring)
 If  $DNAstring_i$ = A, then $T_{i+1} = T_i + 2$
 If  $DNAstring_i$ = G, then $T_{i+1} = T_i + 1$
 If  $DNAstring_i$ = C, then $T_{i+1} = T_i - 1$
 If  $DNAstring_i$ = T, then $T_{i+1} = T_i - 2$
End
```

Human: GTCAAT…AAGAGATTTG

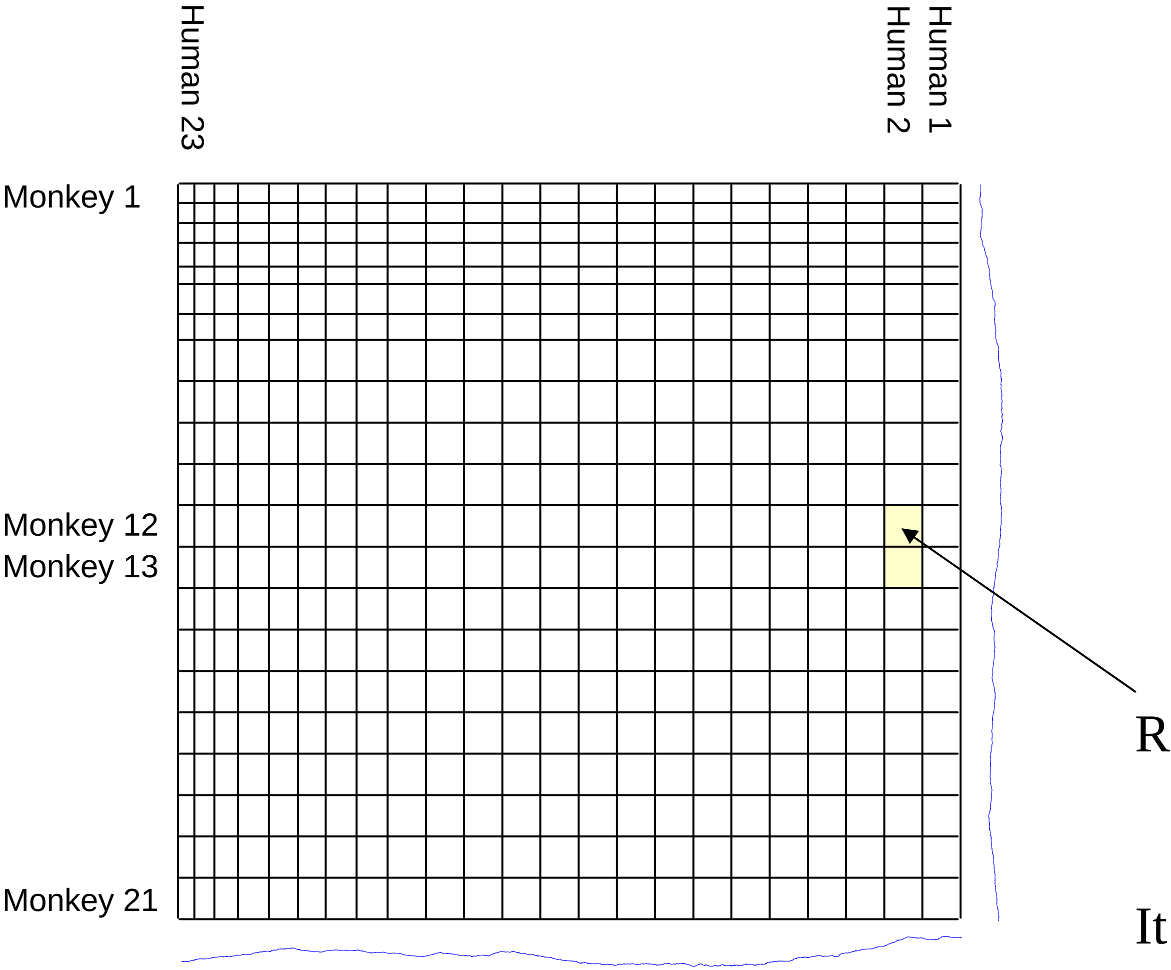




23
chromosomes

21
chromosomes

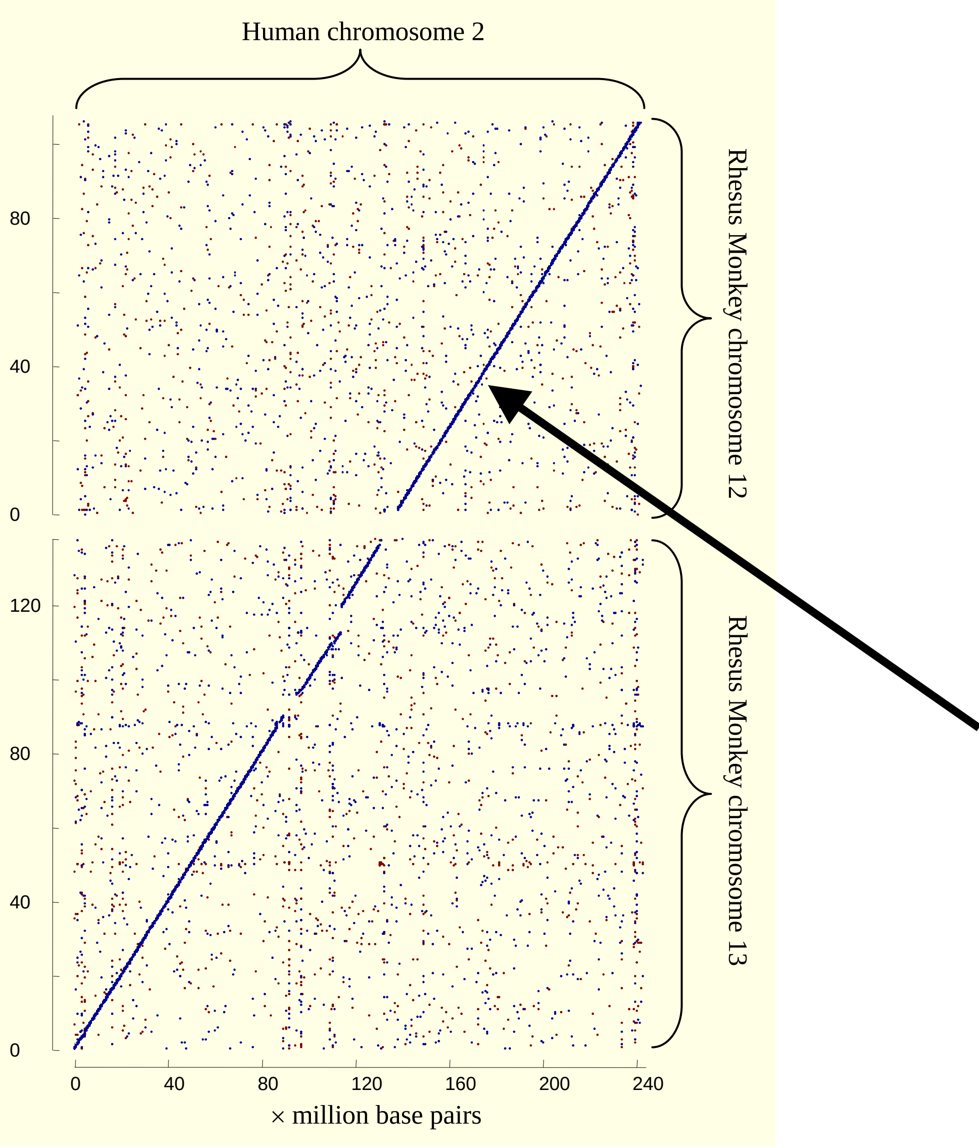
If we motif join the two time series, using subsections of length 1024, where does the first section join?



Right here.

It is hard to see at this scale, lets zoom in…

We have added in a few more points…



How Long Does Motif Discovery Take?

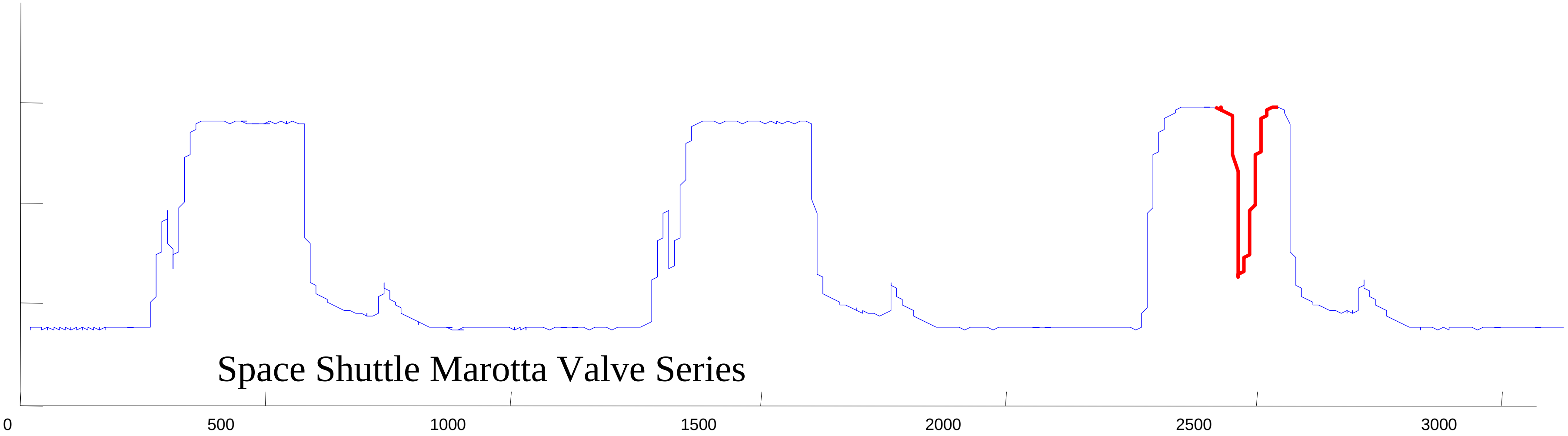
- Naively it takes O(N^2)
- There are a dozen *approximate* algorithms that take O(N) or O(Nlog(N)), with very high constants

- Recently we have invented a fast *exact* algorithm. How fast is it?
  - For the EEG researchers, we reduced their "*over 24 hours*" to 2.1 minutes.
  - Further improvements may be possible

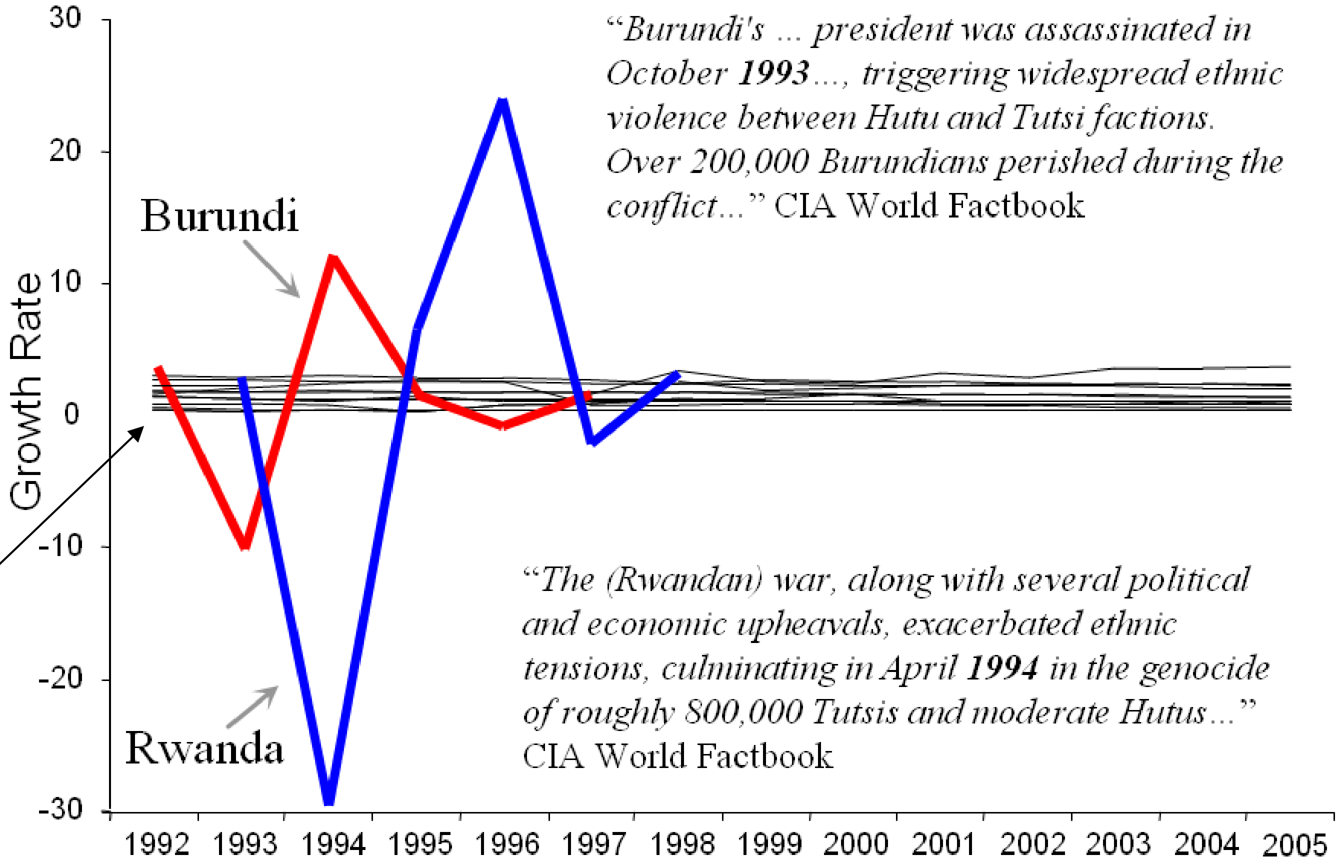
Summary of Motifs

- We now can find exact motifs in very large datasets, truly massive datasets remain a challenge.
- We can use the relative frequency of motifs to monitor a data stream (anomaly detection). We can be alerted by patterns that *don't* exist.
- We could run rule-finding algorithms on discretely labeled "motif stream" This has not been done yet…
- Motifs joins seem very promising…

What are Time Series Discords?



- Population growth data – we studied the growth rate of 206 countries for the last 25 years, looking for the most dramatic 5 year event

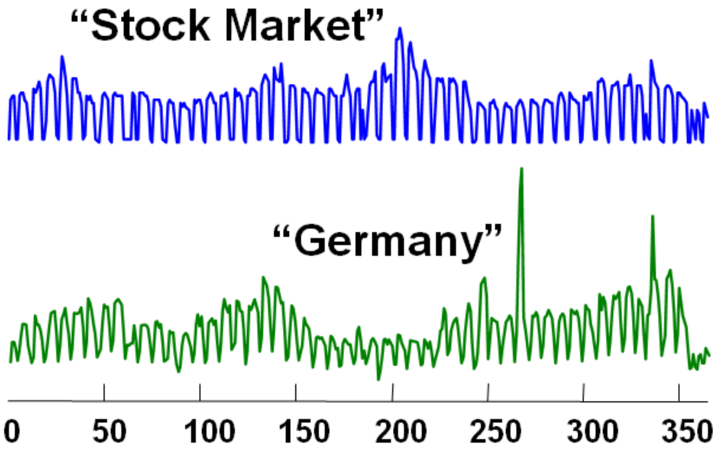


"Burundi's … president was assassinated in October ***1993****…, triggering widespread ethnic violence between Hutu and Tutsi factions. Over 200,000 Burundians perished during the conflict…"* CIA World Factbook

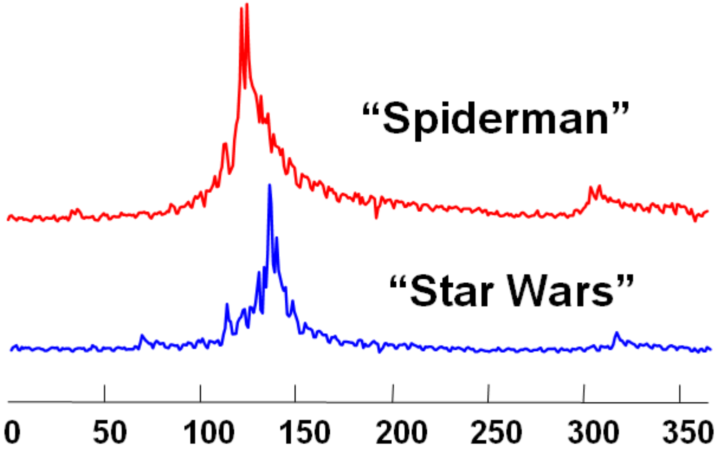
"The (Rwandan) war, along with several political and economic upheavals, exacerbated ethnic tensions, culminating in April ***1994*** *in the genocide of roughly 800,000 Tutsis and moderate Hutus…"* CIA World Factbook

the top 2 discords with a set of 10 representative countries for contrast

MSN web queries made in 2002



patterns dominated by a weekly cycle



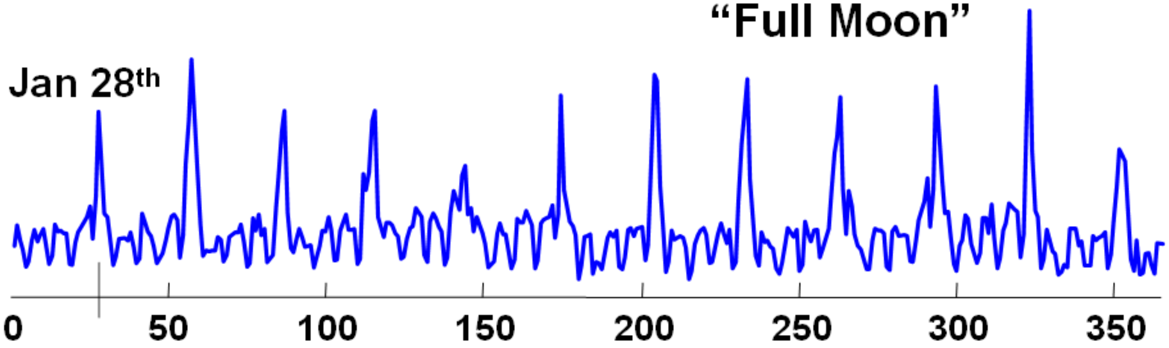
anticipated bursts

- Patterns have daily periodicity (effect of internet at work only?)
- Many patterns have a weekly periodicity (*sports*, *moviephone*)
- Some have monthly periodicity (*insurance plans*, *apartment hunting*)
- Some have yearly periodicity (*Easter*, *Halloween, tour de france* etc)

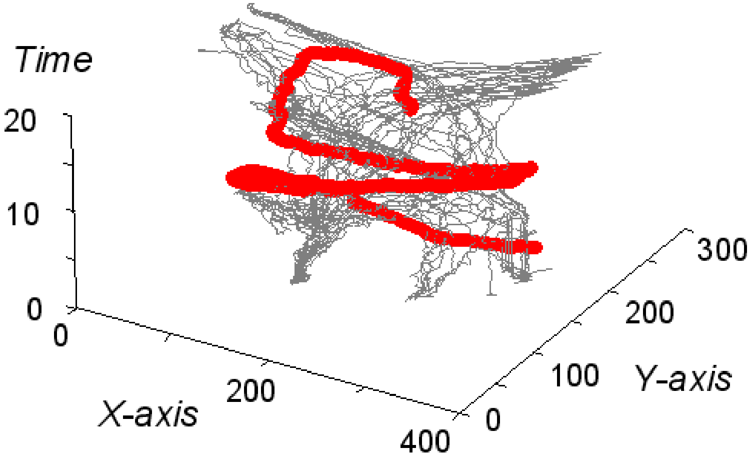
What is the discord (the most unusual web query pattern) in the web logs?

The most significant discord using rotation invariant Euclidean distance

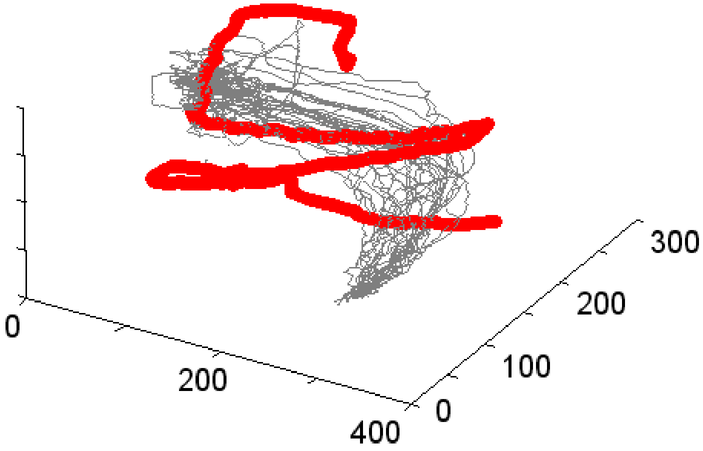
periodicity 29.5 days – the length of a synodic month



- Anomaly detection in video sequences (multivariate data)

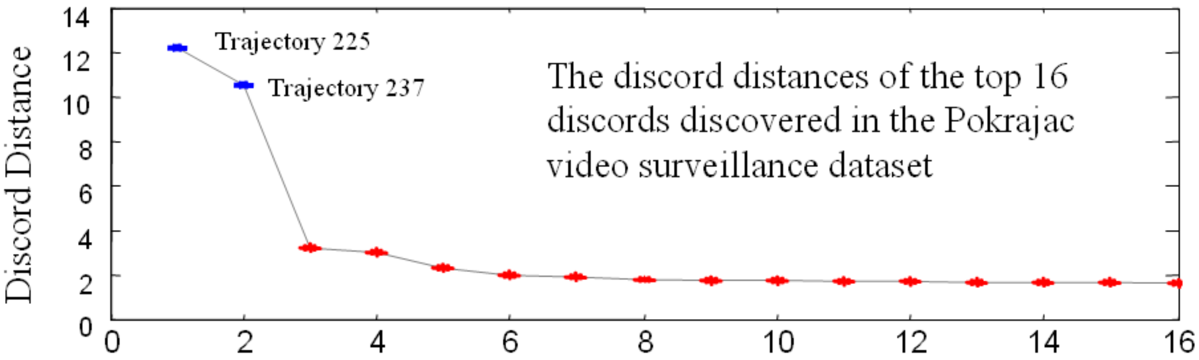


our method achieves 100% accuracy on the planted anomalous trajectories



the top one discord shown with only one of the existing clusters

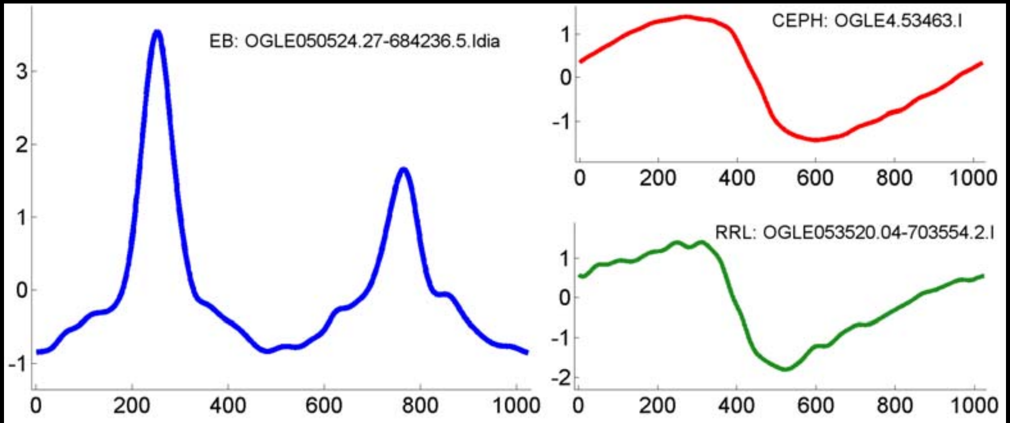
We know there are two anomalies in the dataset, how could we find this automatically?

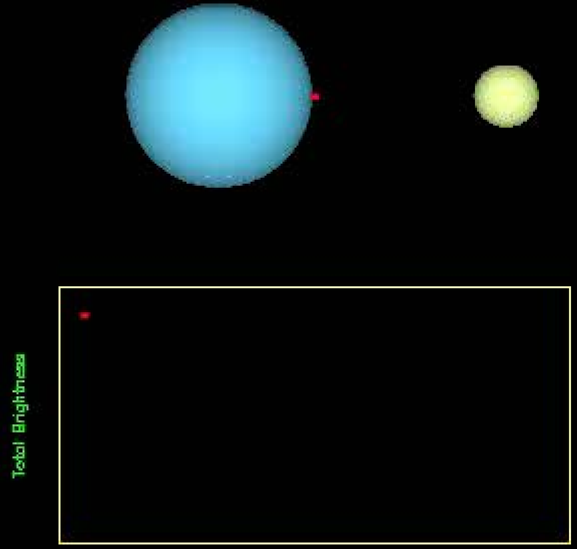


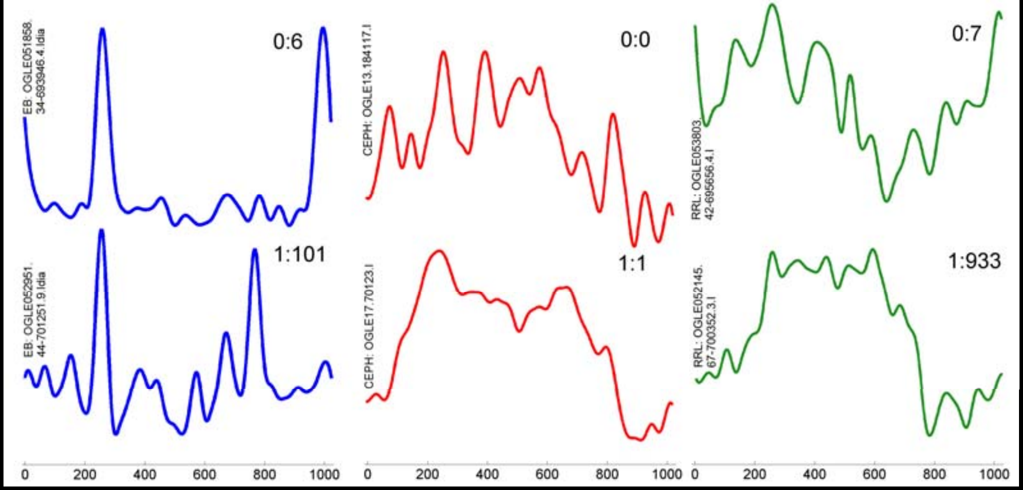
The discord distances of the top 16 discords discovered in the Pokrajac video surveillance dataset

- Star light-curve data from the Optical Gravitational Lensing Experiment (OGLE)

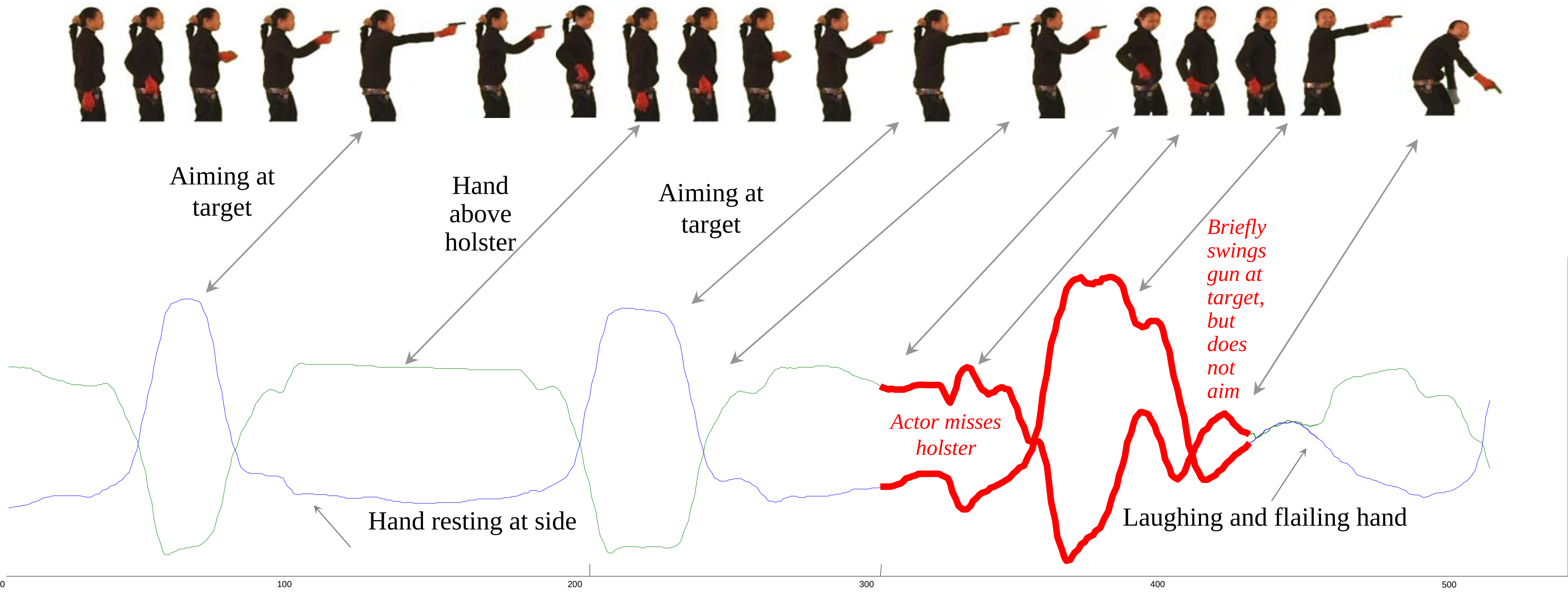
- Three classes of light-curves
  - *Eclipsed binaries*
  - *Cepheids*
  - *RR Lyrae variables*





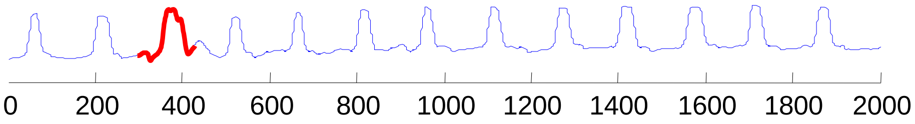


top two discords in each class



The 2D time series was extracted from a video of an actor performing various actions with and without a replica gun. The film strip above illustrates a typical sequence. The two time series measure the X and Y coordinates of the actors right hand. The actor draws a replica gun from a hip mounted holster, aims it at a target, and returns it to the holster. Watching the video we discovered that at about ten seconds into the shoot, the actor misses the holster when returning the gun. An off-camera (inaudible) remark is made, the actor looks toward the video technician, and convulses with laughter. At one point (frame 450), she is literally bent double with laughter.

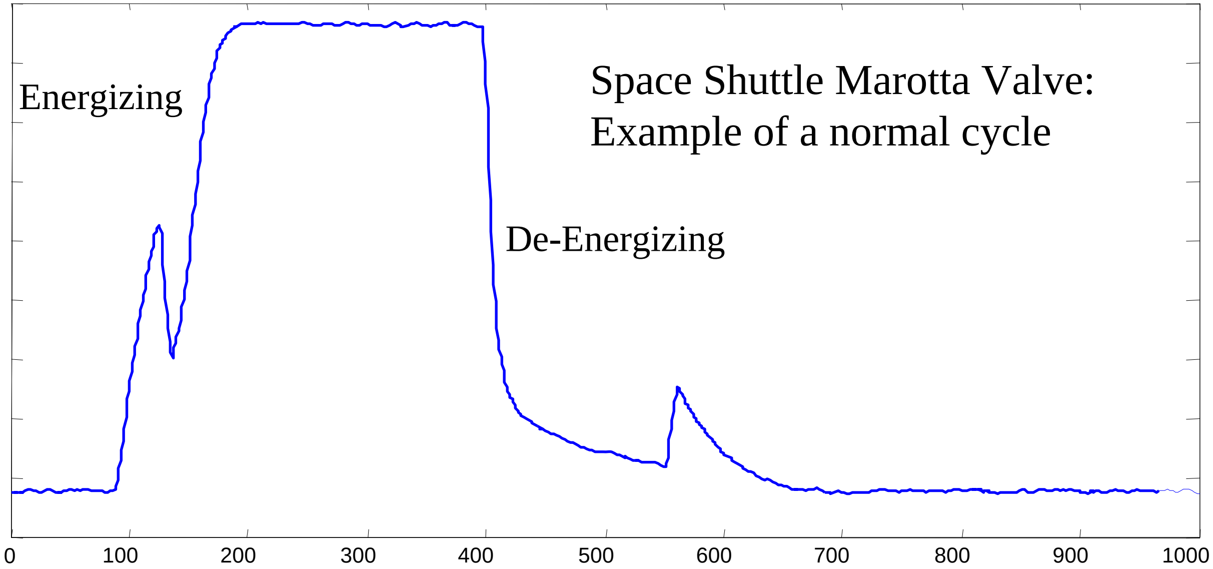
Here is a longer subsection for context.



The next few slides demonstrate utility of discords in finding anomalies in Space Shuttle Marotta Valve time series.

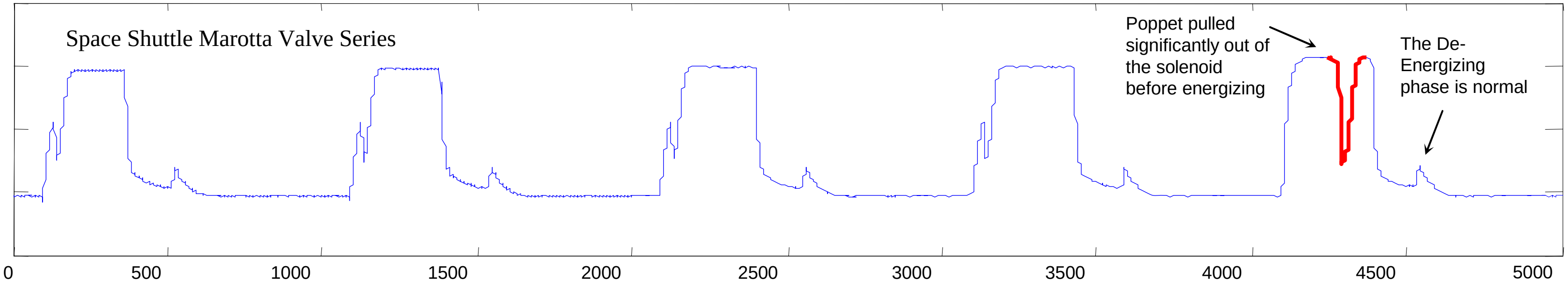
In every case, there are five examples of an Energize/De-Energize cycle. Exactly *one* cycle has been annotated by a domain expert as been abnormal.

We tested on 3 different challenges, of increasing difficulty. Note that all the annotations shown are those of the domain expert.



Test 1: A simple problem

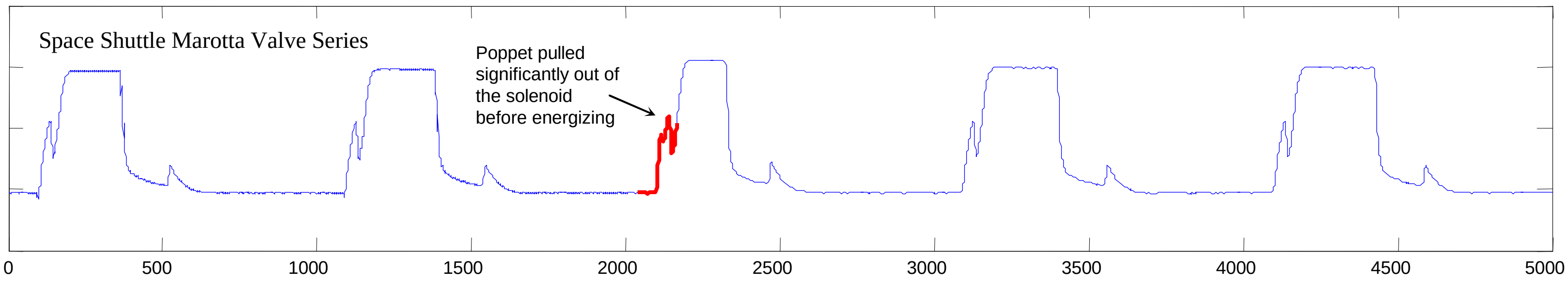
This is dataset TEK16.txt

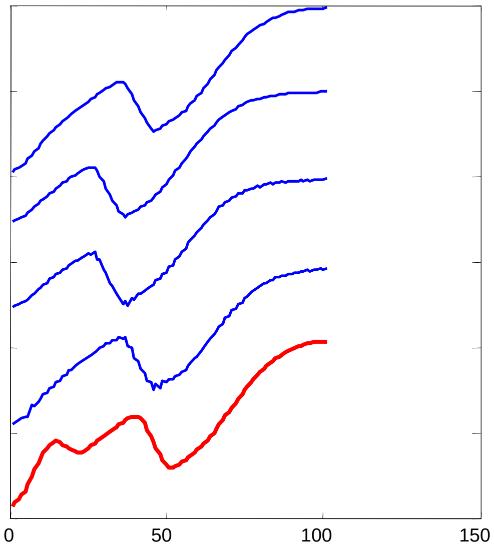


In this case the anomaly is very obvious, and the discord (marked in red) easily finds it.

Test 2: A more subtle problem

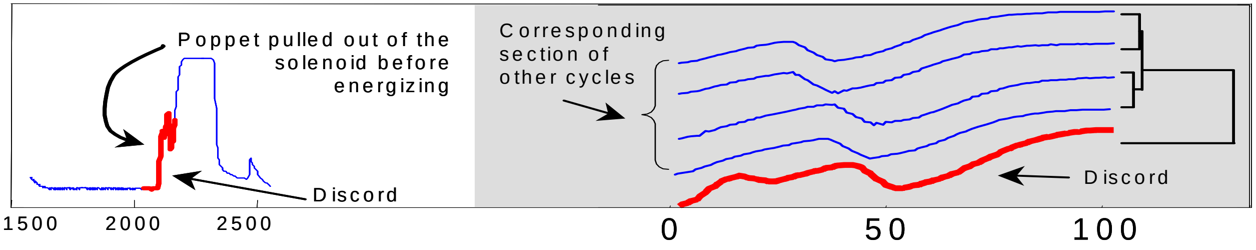
This is dataset TEK17.txt



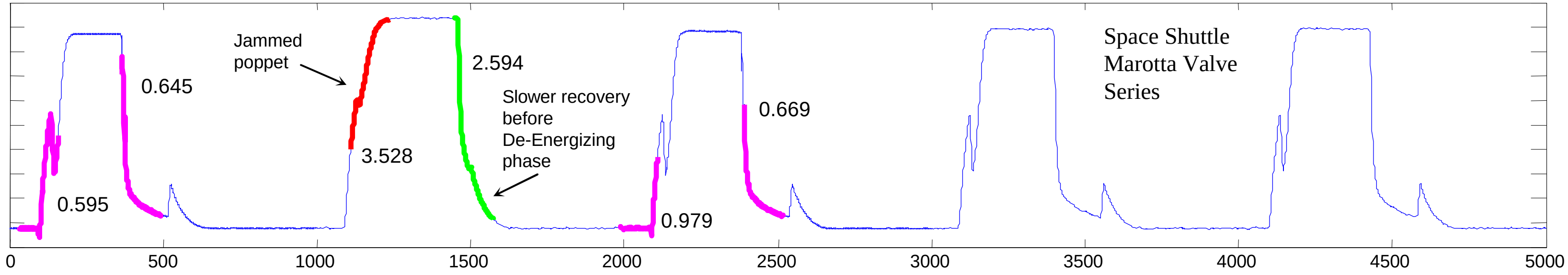


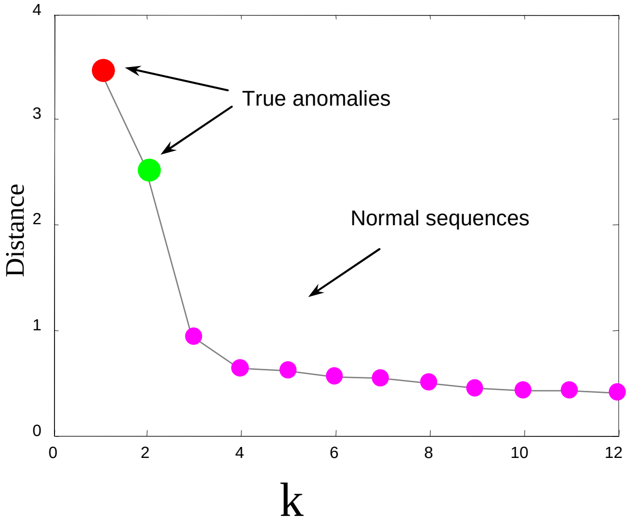
Here the discord (marked in red) easily finds the anomaly marked by the domain expert, but it is not obvious (at this scale) what the anomaly was.

A *zoom-in* of the anomaly, and the 4 corresponding segments from the normal cycle (*left*), explains what the discord discovered. Only the anomalous cycle has a "*double hump*".



Test 3: Finding multiple discords





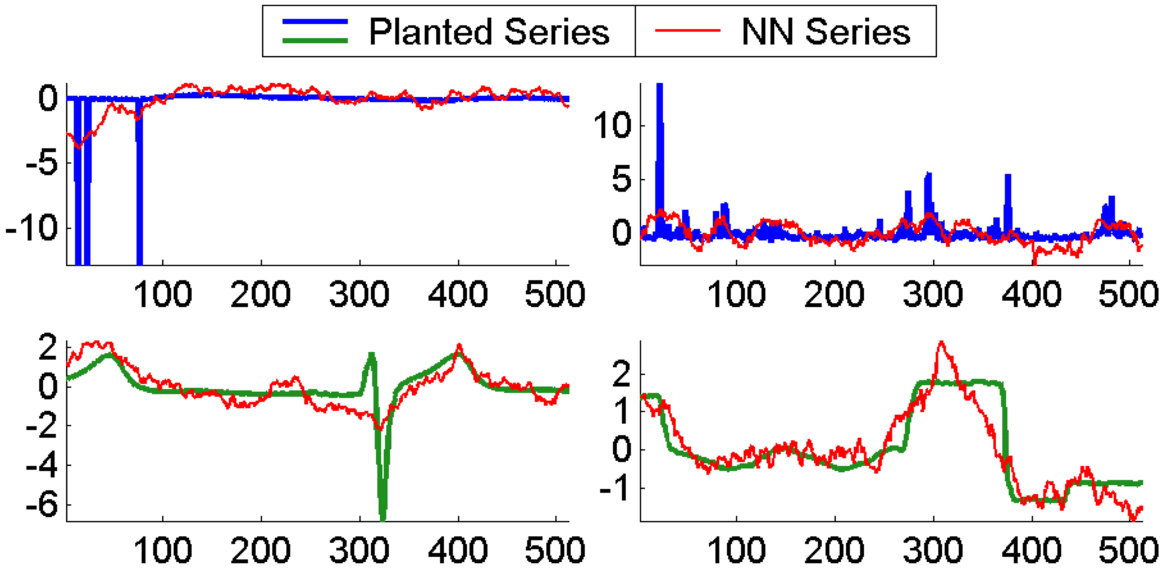
In this example we consider the problem of knowing when an discord is significant.
We found the top 12 discords (only 6 are shown above for clarity). The top 2 correspond to true anomalies, in red we see a missing small peak before the large plateau, and in green we see a slower recovery before the de-energizing phase. The next 4 discords are shown in pink.

If we plot the discords scores against K (*left*) we can see that we could potentially assess the significance of an discord with some kind of “knee finding” algorithm.

Experimental evaluation –
scalability of the disk aware algorithm

- We generated 3 data sets of size up to 0.35Tb of random walk time series

- Six non-random walk time series were planted, we looked for the top 10 discords



two of the planted series (top) were among the top 10 discords

- Time efficiency on the three random walk data sets:

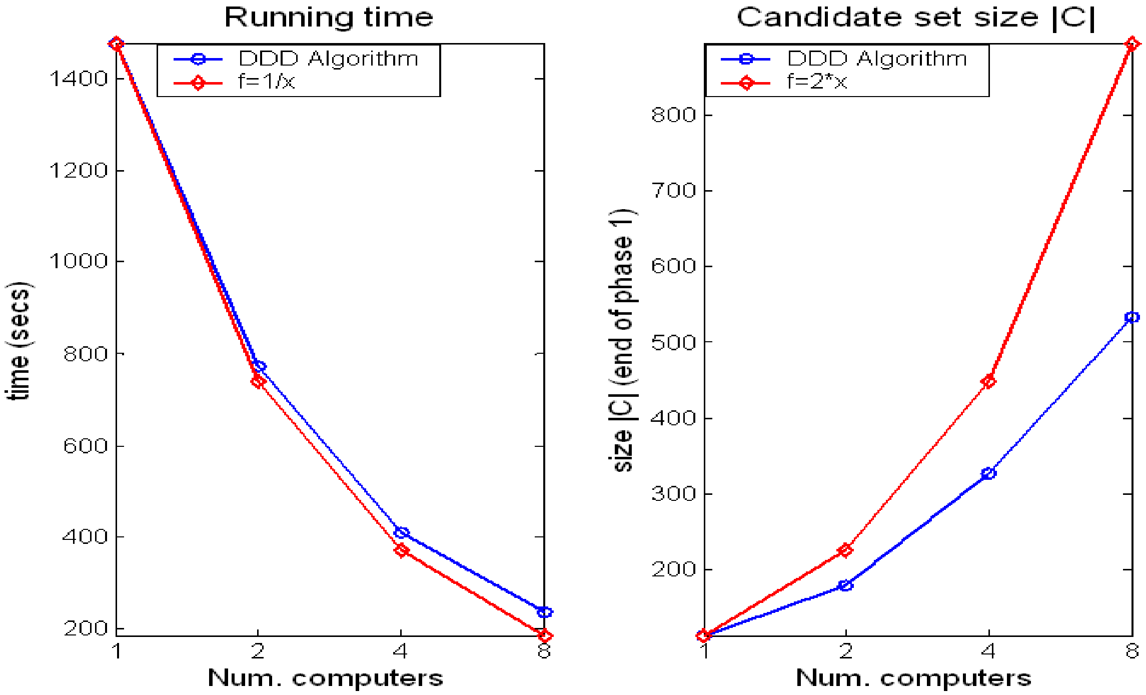
Examples	Disk size	I/O Time	Total time
1 million	3.57 Gb	27min	41min
10 million	35.7 Gb	4h 30min	7h 52min
100 million	0.35 Tb	45h	90h 33min

This is with data on a USB hard drive, and an old laptop

Experimental evaluation –
scalability of the disk aware algorithm

- Parallelizing the algorithm (*dataset: one million random walks*):

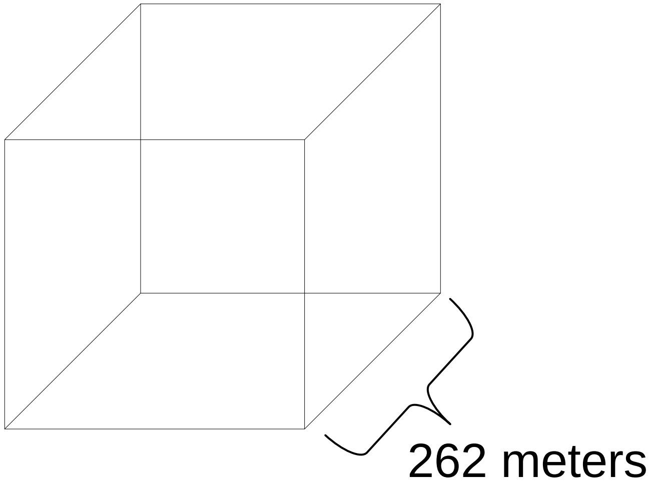
The runtime overhead for 8 computers is approximately 30%.



Needle in a Haystack?

Since the publication of *Don Quixote de la Mancha* in the 17th century, the idiom, "*a needle in a haystack*" has been used to signify a near impossible search.

If each time series in this experiment was represented by a piece of hay the size of a drinking straw, they would form a cube shaped haystack with 262 meter sides.



Summary of Discords

- Works very well empirically in a very wide range of domains.
- With zero or one parameter(s), does not require “tweaking”.
- Very scalable and parallelizable

- Someone needs to port this idea to streaming data…
  - Show me the most unusual thing in the last hour
  - Show me the most unusual thing since this sensor was turned on…

Overall Conclusions

- Motifs, motif joins and discords are very simple but effective tools for understanding massive datasets.
- Parameters are bad, motifs and discords work well because there are essentially no parameters to tweak.
- If you have datasets, problems, internships or money, come talk to me!

Questions?

- Eamonn Keogh
- Computer Science & Engineering Department
- University of California - Riverside
- Riverside, CA 92521
- Phone: (951) 827-2032
- eamonn@cs.ucr.edu

(Twin Freak) Time series discords

- *Most-significant discord* – the subsequence $C_i \in S$ with maximal distance $Dist(C_i, C_j)$ to its N nearest neighbor $C_j \in S$